arXiv:1007.3880v2 [math.ST] 25 Feb 2011

$\sqrt{n}$-CONSISTENT PARAMETER ESTIMATION FOR SYSTEMS OF ORDINARY DIFFERENTIAL EQUATIONS: BYPASSING NUMERICAL INTEGRATION VIA SMOOTHING

SHOTA GUGUSHVILI AND CHRIS A.J. KLAASSEN

Abstract. We consider the problem of parameter estimation for a system of ordinary differential equations from noisy observations on a solution of the system. In case the system is nonlinear, as it typically is in practical applications, an analytic solution to it usually does not exist. Consequently, straightforward estimation methods like the ordinary least squares method depend on repetitive use of numerical integration in order to determine the solution of the system for each of the parameter values considered, and to find subsequently the parameter estimate that minimises the objective function. This induces a huge computational load to such estimation methods. We study the asymptotic consistency of an alternative estimator that is defined as a minimiser of an appropriate distance between a nonparametrically estimated derivative of the solution and the right-hand side of the system applied to a nonparametrically estimated solution. This smooth and match estimator (SME) bypasses numerical integration altogether and reduces the amount of computational time drastically compared to ordinary least squares. Moreover, we show that under suitable regularity conditions this smooth and match estimation procedure leads to a $\sqrt{n}$-consistent estimator of the parameter of interest.

1. Brief Introduction

Many dynamical systems in science and applications are modelled by a d-dimensional system of ordinary differential equations, denoted as

$$\begin{cases} x'(t) = F(x(t),\theta), & t \in [0,1], \\ x(0) = \xi, \end{cases} \tag{1}$$

where θ is the unknown parameter of interest and ξ is the initial condition. With $x_\theta(t)$ the solution vector corresponding to the parameter value θ, we observe

$$Y_{ij} = x_{\theta j}(t_i) + \epsilon_{ij}, \quad i = 1,\ldots,n,\ j = 1,\ldots,d,$$

where the observation times $0 \leq t_1 < \ldots < t_n \leq 1$ are known and the random variables ϵ_{ij} have mean 0 and model measurement errors combined with latent random deviations from the idealised model (1). Under regularity conditions the ordinary least squares estimator

$$\tilde{\theta}_n = \mathrm{argmin}_\eta \sum_{i=1}^{n} \sum_{j=1}^{d} \left(Y_{ij} - x_{\eta j}(t_i)\right)^2 \tag{2}$$

Date: February 28, 2011.

2000 Mathematics Subject Classification. Primary: 62F12, Secondary: 62G08, 62G20.

Key words and phrases. M-estimator; $\sqrt{n}$-consistency; Nonparametric regression; ODE system; Priestley-Chao estimator.

of θ is $\sqrt{n}$-consistent, at least theoretically. For systems (1) that do not have explicit solutions, one typically uses iterative procedures to approximate this ordinary least squares estimator. However, since every iteration in such a procedure involves numerical integration of the system (1) and since the number of iterations is typically very large, in practice it is often extremely difficult if not impossible to compute (2), cf. p. 172 in Voit (2000). Here we present a feasible and computationally much faster method to estimate the parameter θ. To define the estimator of θ we first construct kernel estimators

$$\hat{x}_j(t) = \sum_{i=1}^{n}(t_i - t_{i-1})\frac{1}{b}K\left(\frac{t-t_i}{b}\right)Y_{ij}$$

of $x_{\theta j}$ with K a kernel function and $b = b_n$ a bandwidth. Now, the estimator $\hat{\theta}_n$ of θ is defined as

$$\hat{\theta}_n = \operatorname{argmin}_{\eta} \int_0^1 \| \hat{x}'(t) - F(\hat{x}(t), \eta) \|^2 w(t)\, dt, \tag{3}$$

where $\|\cdot\|$ denotes the usual Euclidean norm and $w(\cdot)$ is a weight function. Related approaches have been suggested in computational biology and numerical analysis literature, see e.g. Bellman and Roth (1971), Voit and Savageau (1982) and Varah (1982).

The main result of this paper is that this smooth and match estimator $\hat{\theta}_n$ is $\sqrt{n}$-consistent under mild regularity conditions. So, asymptotically the SME $\hat{\theta}_n$ is comparable to the ordinary least squares estimator in statistical performance, but it avoids the computationally costly repeated use of numerical integration of (1).

2. Introduction

Let us introduce the contents of this paper in more detail. Systems of ordinary differential equations play a fundamental role in many branches of natural sciences, e.g. mathematical biology, see Edelstein-Keshet (2005), biochemistry, see Voit (2000), or the theory of chemical reaction networks in general, see for instance Feinberg (1979) and Sontag (2001). Such systems usually depend on parameters, which in practice are often only approximately known, or are plainly unknown. Knowledge of these parameters is critical for the study of the dynamical system or process that the system of ordinary differential equations describes. Since these parameters usually cannot be measured directly, they have to be inferred from, as a rule, noisy measurements of various quantities associated with the process under study. More formally, in this paper we consider the following setting: let, as in (1),

$$\begin{cases} x'(t) = F(x(t), \theta), & t \in [0,1], \\ x(0) = \xi, \end{cases} \tag{4}$$

be a system of autonomous differential equations depending on a vector of real-valued parameters. Here $x(t) = (x_1(t), \ldots, x_d(t))^T$ is a d-dimensional state variable, $\theta = (\theta_1, \ldots, \theta_p)^T$ denotes a p-dimensional parameter, while the column d-vector $x(0) = \xi$ defines the initial condition. Whether the latter is known or unknown, is not relevant in the present context, as long as it stays fixed. Denote a solution to (4) corresponding to parameter value θ by $x_\theta(t) = (x_{\theta 1}(t), \ldots, x_{\theta d}(t))^T$. Suppose that at known time instances $0 \leq t_1 < \ldots < t_n \leq 1$ noisy observations

$$Y_{ij} = x_{\theta j}(t_i) + \epsilon_{ij}, \quad i = 1, \ldots, n,\ j = 1, \ldots, d, \tag{5}$$

on the solution x_θ are available. The random variables ϵ_{ij} model measurement errors, but they might also contain latent random deviations from the idealized model (1). Such random deviations are often seen in real-world applications. Based on these observations, the goal is to infer the value of θ, the parameter of interest.

The standard approach to estimation of θ is based on the least squares method (the least squares method is credited to Gauß and Legendre, see Stigler (1981)), see e.g. Hemker (1972) and Stortelder (1996). The least squares estimator is defined as a minimiser of the sum of squares, i.e.

$$\tilde{\theta}_n = \operatorname{argmin}_\eta R_n(\eta) = \operatorname{argmin}_\eta \sum_{i=1}^{n} \sum_{j=1}^{d} (Y_{ij} - x_{\eta j}(t_i))^2.$$

If the measurement errors are Gaussian, then $\tilde{\theta}_n$ coincides with the maximum likelihood estimator and is asymptotically efficient. Since the differential equations setting is covered by the general theory of nonlinear least squares, theoretical results available for the latter apply also in the differential equations setting and we refer e.g. to Jennrich (1969) and Wu (1981), or more generally to van de Geer (1990), van de Geer and Wegkamp (1996), and Pollard and Radchenko (2006) for a thorough treatment of the asymptotics of the nonlinear least squares estimator. The paper that explicitly deals with the ordinary differential equations setting is Xue et al. (2010). Despite its appealing theoretical properties, in practice the performance of the least squares method can dramatically degrade if (4) is a nonlinear high-dimensional system and if θ is high-dimensional. In such a case we have to face a nonlinear optimisation problem (quite often with many local minima) and search for a global minimum of the least squares criterion function R_n in a high-dimensional parameter space. The search process is most often done via gradient-based methods, e.g. the Levenberg-Marquardt method, see Marquardt (1963), or via random search algorithms, see Section 4.5.2 in Voit (2000) for a literature overview. Since nonlinear systems in general do not have solutions in closed form, use of numerical integration within a gradient-based search method and serious computational time associated with it seem to be inevitable. For instance, a relatively simple example of a four-dimensional system considered in Appendix 1 of Voit and Almeida (2004) demonstrates that the need to repeat numerical integration multiple times might increase the computational time for numerical integration up to 95% of the total computational time required for a gradient based optimisation method. Likewise, random search algorithms are also very costly computationally and in general, computational time will typically be a problem for any optimisation algorithm that relies on numerical integration of any relatively realistic nonlinear system of ordinary differential equations, cf. p. 172 in Voit (2000). One example is furnished by Kikuchi et al. (2003), where a system that consists of five differential equations and contains sixty parameters and that describes a simple gene regulatory network from Hlavacek and Savageau (1996) is considered. The optimisation algorithm (a genetic algorithm) was run for seven loops each lasting for about ten hours on the AIST CBRC Magi Cluster with 1040 CPUs (Pentium III 933 MHz)[1]. This amounted to a total of ca. 70,000 CPU hours. The authors also remarked that the gradient-based search algorithm would not be feasible in their setting at all. The problems become aggravated for systems of ordinary differential equations

[1] See http://www.cbrc.jp/magi for the cluster specifications.

that exhibit stiff behaviour, i.e. systems that contain both 'slow' and 'fast' variables and that are difficult to integrate via explicit numerical integration schemes, see e.g. Hairer and Wanner (1996) for a comprehensive treatment of methods of solving numerically stiff systems. Even if a system is not stiff for the true parameter value θ, during the numerical optimisation procedure one might pass the vicinity of parameters for which the system is stiff, which will necessarily slow down the optimisation process.

The Bayesian approach to estimation of θ, see e.g. Gelman et al. (1996) and Girolami (2008), encounters similar huge computational problems. In the Bayesian approach one puts a prior on the parameter θ and then obtains the posterior via Bayes' formula. The posterior contains all the information required in the Bayesian paradigm and can be used to compute e.g. point estimates of θ or Bayesian credible intervals. If θ is high-dimensional, the posterior will typically not be manageable by numerical integration and one will have to resort to Markov Chain Monte Carlo (MCMC) methods. However, sampling from the posterior distribution for θ via MCMC necessitates at each step numerical integration of the system (4), in case the latter does not have a closed form solution. Computational time might thus become a problem in this case as well. Also, since in general the likelihood surface will have a complex shape with many local optima, ripples, and ridges, see e.g. Girolami (2008) for an example, serious convergence problems might arise for MCMC samplers.

Yet another point is that in practice both the least squares method and the Bayesian approach require good initial guesses of the parameter values. If these are not available, then both approaches might have problems with convergence to the true parameter value within a reasonable amount of time.

Over the years a number of improvements upon the classical methods to compute the least squares estimator have been proposed in the literature. In particular, the multiple shooting method of Bock (1983) and the interior-point or barrier method for large-scale nonlinear programming as in Wächter and Biegler (2006) have proved to be quite successful. These two approaches tend to be much more stable than classical gradient-based methods, have a better chance to converge even from poor initial guesses of parameters, and in general require a far less number of iterations until convergence is achieved. However, they still require a nontrivial amount of computational power.

A general overview of the typical difficulties in parameter estimation for systems of ordinary differential equations is given in Ramsay et al. (2007), to which we refer for more details. For a recent overview of typical approaches to parameter estimation for systems of ordinary differential equations in biochemistry and associated challenges see e.g. Chou and Voit (2009).

To evade difficulties associated with the least squares method, or more precisely with numerical integration that it usually requires, a two-step method was proposed in Bellman and Roth (1971) and Varah (1982). In the first step the solution x_θ of (4) is estimated by considering estimation of the individual components $x_{\theta 1},\ldots,x_{\theta d}$ as nonparametric regression problems and by using the regression spline method for estimation of these components. The derivatives of $x_{\theta 1},\ldots,x_{\theta d}$ are also estimated from the data by differentiating the estimators of $x_{\theta 1},\ldots,x_{\theta d}$ with respect to time t. Thus no numerical integration of the system (4) is needed. In the second step the obtained estimate of x_θ and its derivative x'_θ are plugged into (4) and an estimator of θ is defined as a minimiser in θ of an appropriate distance between the estimated left- and righthand sides of (4) as e.g. in (3). Since this estimator of θ results from a minimisation procedure, it is an M-estimator, see e.g. the classical monograph Huber (1981), or Chapter 7 of Bickel et al. (1998), Chapter 5 of van der Vaart (1998), and Chapter 3.2 of Wellner and van der Vaart (1996) for a more modern exposition of the theory of M-estimators. For an approach to estimation of θ related to Bellman and Roth (1971) and Varah (1982) see also Voit and Savageau (1982), as well as Voit and Almeida (2004), where a practical implementation based on neural networks is studied. The intuitive idea behind the use of this two-step estimator is clear: among all functions defined on $[0, 1]$, any reasonably defined distance between the left- and righthand side of (4) is minimal (namely, it is zero) for the solution x_θ of (4) and the true parameter value θ. For estimates close enough in an appropriate sense to the solution x_θ, the minimisation procedure will produce a minimiser close to the true parameter value, provided certain identifiability and continuity conditions hold. This intuitive idea was exploited in Brunel (2008), where a more general setting than the one in Bellman and Roth (1971) and Varah (1982) was considered. Another paper in the same spirit as Bellman and Roth (1971) and Varah (1982) is Liang and Wu (2008).

This two-step approach will typically lead to considerable savings in computational time, as unlike the straightforward least squares estimator, in its first step it just requires finding nonparametric estimates of x_θ and x'_θ, for which fast and numerically reliable recipes are available, whereas the gradient-based least squares method will still rely on successive numerical integrations of (4) for different parameter values θ in order to find a global minimiser minimising the least squares criterion function. We refer to Voit and Almeida (2004) for a particular example demonstrating gains in the computational time achieved by the two-step estimator in comparison to the ordinary least squares estimator. When the righthand side F of (4) is linear in $\theta_1, \dots, \theta_p$ and $d = 1$, further simplifications will occur in the second step of the two-step estimation procedure, as one will essentially only have to face a weighted linear regression problem then. This is unlike the least squares approach, which cannot exploit linearity of F in $\theta_1, \dots, \theta_p$. However, we would also like to stress the fact that the two-step estimator does not necessarily have to be considered a competitor of either the least squares or the Bayesian approach. Indeed, since in practice both of these approaches require good initial guesses for parameter values, these can be supplied by the two-step estimator. In this sense the proposed two-step estimation approach can be thought of as complementing both the least squares and the Bayesian approaches. Moreover, an additional modified Newton-Raphson step suffices to arrive at an estimator that is asymptotically equivalent to the exact ordinary least squares estimator, as will be shown elsewhere.

A certain limitation of the two-step approach is that it requires that measurements on all state variables $x_{\theta j}, j = 1, \dots, d$ are available. The latter is not always the case in practical applications. In some cases the unobserved variables can be eliminated by transforming the first order system into a higher order one and next applying a generalisation of our smooth and match method to this higher order system. This approach should yield a consistent estimator. One might also formally perform the least squares procedure in such a case. However, without stringent assumptions on the system it is far from clear that this leads to a consistent estimator.

Our goal in the present work is to undertake a rigorous study of the asymptotics of a two-step estimator of θ. Our exposition is similar to that in Brunel (2008) to some degree, but one of the differences is that instead of regression spline estimators we use kernel-type estimators for estimation of x_θ and x'_θ.[2] The conditions are also different. We hope that our contribution will motivate further research into the interesting topic of parameter estimation for systems of ordinary differential equations.

There exists an alternative approach to the ones described here, which also employs nonparametric smoothing, see Ramsay et al. (2007). For information on its asymptotic properties we refer to Qi and Zhao (2010). For nonlinear systems this appproach will typically reduce to one of the realisations of the ordinary least squares method, e.g. Newton-Raphson algorithm, where however numerical integration of (4) will be replaced by approximation of the solution of the system (4) by an appropriately chosen element of some finite-dimensional function space. This seems to reduce considerably the computational load in comparison to the gradient-based optimisation methods which employ numerical integration of (4). However, it still appears to be computationally more intense than the two-step approach advocated in the present work.

We conclude the discussion in this section by noting that when modelling various processes, some authors prefer not to specify the righthand side of (4) explicitly (the latter amounts to explicit specification of the $F(\cdot,\cdot)$ in (4)), but simply assume that the righthand side of (4) is some unknown function of x, i.e. is given by $F(x(t))$ with F unknown, and proceed to its estimation via nonparametric methods, see e.g. Ellner et al. (2002). This has an advantage of safeguarding against possible model misspecification. However, the question whether one has or has not to specify F explicitly appears to us to be more of a philosophical nature and boils down to a discussion on the use of parametric or nonparametric models, i.e. whether one has strong enough reasons to believe that the process under study can be described by a model as in (4) with F known or not. We do not address this question here, because an answer to it obviously depends on the process under study and varies from case to case. For a related discussion see Hooker (2009).

The rest of the paper is organised as follows: in the next section we will detail the approach that we use and present its theoretical properties. In particular, we will show that under appropriate conditions our two-step approach leads to a consistent estimator with a $\sqrt{n}$ convergence rate, which is the best possible rate in regular parametric models[3]. Section 4 contains a discussion on the obtained results together with simulation examples. The proofs of the main results are relegated to Section 5, while the Appendices contain some auxiliary statements.

3. Results

First of all, we point out that in the present study we will be concerned with the asymptotic behaviour of an appropriate two-step estimator of θ under a suitable sampling scheme. We will primarily be interested in intuitively understanding the behaviour of a relatively simple estimator of θ, as well as in a clear presentation of

[2] The proofs of the main results in Brunel (2008) are incomplete and the main theorems require further conditions in order to hold.

[3] It is claimed in Liang and Wu (2008) that their two-step estimation procedure leads to a faster rate than $\sqrt{n}$, which is impossible. Indeed, their Theorem 2 and its proof are incorrect.

the obtained results and the proofs. Consequently, the stated conditions will not always be minimal and can typically be relaxed at appropriate places.

We first define the sampling scheme.

Condition 1. *The observation times $0 \leq t_1 < \ldots < t_n \leq 1$ are deterministic and known and there exists a constant $c_0 \geq 1$, such that for all n*

$$\max_{2 \leq i \leq n} |t_i - t_{i-1}| \leq \frac{c_0}{n}$$

holds. Furthermore, there exists a constant $c_1 \geq 1$, such that for any interval $A \subseteq [0,1]$ of length $|A|$ and all $n \geq 1$ the inequality

$$\frac{1}{n} \sum_{i=1}^{n} 1_{[t_i \in A]} \leq c_1 \max\left(|A|, \frac{1}{n} \right)$$

holds.

Hence, we observe the solution of the system (4) on the interval $[0,1]$. Instead of $[0,1]$ we could have taken any other bounded interval. Conditions on $t_1, \ldots, t_n$ as in Condition 1 are typical in nonparametric regression, see e.g. Gasser and Müller (1984) and Section 1.7 in Tsybakov (2009), and they imply that $t_1, \ldots, t_n$ are distributed over $[0,1]$ in a sufficiently uniform manner. The most important example in which Condition 1 is satisfied, is when the observations are spaced equidistantly over $[0,1]$, i.e. when $t_j = j/n$ for $j = 1, \ldots, d$. In this case one may take $c_0 = c_1 = 2$. Notice that we do not necessarily assume that the initial condition $x(0) = \xi$ is measured or is known. If it is, then it is incorporated into the observations and is used in the first step of the two-step estimation procedure.

Condition 2. *The random variables ϵ_{ij} , $i = 1, \ldots, n$, $j = 1, \ldots, d$, from (5) are independent and are normally distributed with mean zero and finite variance σ_j^2.*

This assumption of Gaussianity of the ϵ_{ij}'s may be dropped in various ways, as we will see below; see the note after Proposition 1 and Appendix B.

We next state a condition on the parameter set.

Condition 3. *The parameter set Θ is a compact subset of $\mathbb{R}^p$.*

Compactness of Θ allows one to put relatively weak conditions on the structure of the system (4), i.e. the function F.

Just as the least squares method, see e.g. Jennrich (1969), our smooth and match approach also requires some regularity of the solutions of (4). In what follows, a derivative of any function f with respect to the variable y will be denoted by f'_y. For the second derivative of f with respect to y we will use the notation f''_{yy} with a similar convention for mixed derivatives. An integral of a vector- or matrix-valued function will be understood componentwise.

Condition 4. *The following conditions hold:*

(i) *the mapping $F : \mathbb{R}^d \times \Theta \to \mathbb{R}^d$ from (4) is such that its second derivatives $F''_{\theta\theta}, F''_{\theta x}, F''_{xx}$ are continuous;*

(ii) *for all parameter values $\theta \in \Theta$, the solution x_θ of (4) is defined on the interval $[0,1]$;*

(iii) *for all parameter values $\theta \in \Theta$, the solution x_θ of (4) is unique on $[0,1]$;*

(iv) *for all parameter values $\theta \in \Theta$, the solution x_θ of (4) is a C^α function of t on the interval $[0,1]$ for some positive integer α.*

Observe that Condition 4 (i) implies existence and uniqueness of the solution of (4) in some neighbourhood of 0. However, we want the existence and uniqueness to hold on the whole interval $[0,1]$ and therefore a priori require (ii) and (iii). Furthermore, $\alpha \geq 2$ in (iv) is required when establishing appropriate asymptotic properties of nonparametric estimators of the solution x_θ and its derivative, while $\alpha \geq 3$ is needed in Propositions 3 and 4, and $\alpha \geq 4$ in Theorem 1, respectively. Notice that for every θ the solution x_θ is of class C^α in t in a neighbourhood of 0, provided for a given θ the function F is of class C^α in its first argument. However, we want this to hold on the whole interval $[0,1]$ and therefore require (iv). Since in the theory of chemical reaction networks, see for instance Sontag (2001), the components of F are usually polynomial or rational functions of $x_1,\ldots,x_d$ and $\theta_1,\ldots,\theta_p$, the solution of (4) will be smooth enough in many examples and $\alpha \geq 4$ is satisfied in a large number of practical examples. For the above-mentioned facts from the theory of ordinary differential equations see e.g. Chapter 2 in Arnold (1973). Also notice that the condition on F in Liang and Wu (2008), see Assumption C on p. 1573, puts severe restrictions on F and excludes e.g. quadratic nonlinearities of F in $x_1,\ldots,x_d$. This, of course, has to be avoided.

Recall that our observations are $Y_{ij} = x_{\theta j}(t_i) + \epsilon_{ij}$ for $i = 1,\ldots,n, j = 1,\ldots,d$. We propose the following nonparametric estimator for $x_{\theta j}$,

$$\hat{x}_j(t) = \sum_{i=1}^{n}(t_i - t_{i-1})\frac{1}{b}K\left(\frac{t-t_i}{b}\right)Y_{ij}, \tag{6}$$

where K is a kernel function, while the number $b = b_n > 0$ denotes a bandwidth that we take to depend on the sample size n in such a way that $b_n \to 0$ as $n \to \infty$. In line with a traditional convention in kernel estimation theory, we will suppress the dependence of b_n on n in our notation, since no confusion will arise. When the t_i's are equispaced, the estimator (6) can in essence be obtained by modifying the Nadaraya-Watson regression estimator, cf. p. 34 in Tsybakov (2009). It is usually called the Priestley-Chao estimator after the authors who first proposed it in Priestley and Chao (1972). As far as an estimator of $x'_{\theta j}(t)$ is concerned, we define it as the derivative of $\hat{x}_j(t)$ with respect to t, choosing K as a differentiable function. Notice that the bandwidth b plays a role of regularisation parameter: too small a bandwidth results in an estimator with small bias, but large variance, while too large a bandwidth results in an estimator with small variance, but large bias, see e.g. pp. 7–8 and 32 in Tsybakov (2009) for a relevant discussion. In principle one could use different bandwidth sequences for estimation of x_j for different j's, but as can be seen from the proofs in Section 5, asymptotically this will not make a difference for an estimator of θ. A similar remark applies to the use of different bandwidths for estimation of $x_{\theta j}$ and its derivative $x'_{\theta j}$. Arguably, the estimator (6) is simple and there exist other estimators that may outperform it in certain respects in practice. However, as we will show later on, even such a simple estimator leads to a $\sqrt{n}$-consistent estimator of θ.

Theoretical properties of the Priestley-Chao estimator were studied in Benedetti (1977), Priestley and Chao (1972), and Schuster and Yakowitz (1979). However, the first two papers do not cover its convergence in the L_∞ (supremum) norm, while the third one does not do it in the form required in the present work. Since this is needed in the sequel, we will supply the required statement, see Proposition 1 below.

To put things in a somewhat more general context than the one in our differential equations setting, consider the following regression model:

$$\begin{aligned} &Y_i = \mu(t_i) + \epsilon_i, \quad i = 1, \ldots, n, \\ &t_1, \ldots, t_n \text{ satisfy Condition 1 }, \\ &\epsilon_1, \ldots, \epsilon_n \text{ are i.i.d. Gaussian with } \mathbb{E}\left[\epsilon_i\right] = 0 \text{ and } \mathbb{V}\mathrm{ar}\left[\epsilon_i\right] = \sigma^2 > 0. \end{aligned} \tag{7}$$

Our goal is to estimate the regression function μ and its derivative μ'. The estimator of μ will be given by an expression similar to (6), namely

$$\hat{\mu}_n(t) = \sum_{i=1}^{n} (t_i - t_{i-1}) \frac{1}{b} K\left(\frac{t - t_i}{b}\right) Y_i, \tag{8}$$

while an estimator of μ' will be given by $\hat{\mu}_n'$. We postulate the following condition on the kernel K for some strictly positive integer α.

Condition 5. *The kernel K is symmetric and twice continuously differentiable, it has support within $[-1,1]$, and it satisfies the integrability conditions: $\int_{-1}^{1} K(u)du = 1$ and $\int_{-1}^{1} u^{\ell} K(u)du = 0$ for $\ell = 1, \ldots, \alpha - 1$. If $\alpha = 1$, only the first of the two integrability conditions is required.*

The following proposition holds.

Proposition 1. *Suppose the regression model (7) is given and Condition 5 holds. Fix a number δ, such that $0 < \delta < 1/2$.*

(i) If μ is $\alpha \geq 1$ times continuously differentiable and $b \to 0$ as $n \to \infty$, then

$$\sup_{t \in [\delta, 1-\delta]} |\hat{\mu}_n(t) - \mu(t)| = O_P\left(b^{\alpha} + \frac{1}{nb^2} + \sqrt{\frac{\log n}{nb}}\right). \tag{9}$$

(ii) If μ is $\alpha \geq 2$ times continuously differentiable and $b \to 0$ as $n \to \infty$, then

$$\sup_{t \in [\delta, 1-\delta]} |\hat{\mu}_n'(t) - \mu'(t)| = O_P\left(b^{\alpha-1} + \frac{1}{nb^3} + \sqrt{\frac{\log n}{nb^3}}\right) \tag{10}$$

is valid. In particular, $\hat{\mu}_n$ and $\hat{\mu}_n'$ are consistent on $[\delta, 1-\delta]$, if $nb^3/\log n \to \infty$ holds additionally.

Gaussianity of the ϵ_i's allows one to prove (9) and (10) by relatively elementary means. This assumption can be modified in various ways, for instance by assuming that the ϵ_i's are bounded, and we state and prove the corresponding modification of Proposition 1 in Appendix B, see Proposition 5. In general, normality of the measurement errors is a standard assumption in parameter estimation for systems of ordinary differential equations, see e.g. Girolami (2008), Hemker (1972), and Ramsay et al. (2007).

The following corollary is immediate from Proposition 1.

Corollary 1. *Let α be the same as in Condition 4. Under Conditions 1–5 we have for the estimator $\hat{x}_j$*

$$\sup_{t \in [\delta, 1-\delta]} |\hat{x}_j(t) - x_{\theta j}(t)| = O_P\left(b^{\alpha} + \frac{1}{nb^2} + \sqrt{\frac{\log n}{nb}}\right) \tag{11}$$

and

$$\sup_{t\in[\delta,1-\delta]} |\hat{x}_j'(t) - x_{\theta j}'(t)| = O_P\left(b^{\alpha-1} + \frac{1}{nb^3} + \sqrt{\frac{\log n}{nb^3}}\right), \tag{12}$$

provided $\alpha \geq 2$ and $b \to 0$ as $n \to \infty$. In particular, $\hat{x}_j$ and $\hat{x}_j'$ are consistent, if $nb^3/\log n \to \infty$ holds additionally.

In the proof of Proposition 2 we will apply the continuous mapping theorem in order to prove convergence in probability of certain integrals of F and its derivatives with $\hat{x}_j$'s plugged in. This is where Corollary 1 is used.

Now that we have consistent (in an appropriate sense) estimators of $x_{\theta j}$ and $x_{\theta j}'$, from the smoothing step we can move to the matching step in the construction of our smooth and match estimator of θ. In particular, we define the estimator $\hat{\theta}_n$ of θ as

$$\begin{aligned} \hat{\theta}_n &= \operatorname{argmin}_{\eta\in\Theta} \int_0^1 \|\hat{x}'(t) - F(\hat{x}(t),\eta)\|^2 w(t)dt \\ &= \operatorname{argmin}_{\eta\in\Theta} M_{n,w}(\eta), \end{aligned} \tag{13}$$

where $\|\cdot\|$ denotes the usual Euclidean norm and w is a weight function. We will refer to $M_{n,w}(\eta)$ as a (random) criterion function. Since Θ is compact and $M_{n,w}$ under our conditions is continuous in η, the minimiser $\hat{\theta}_n$ always exists. The fact that $\hat{\theta}_n$ is a measurable function of the observations Y_{ij} follows from Lemma 2 of Jennrich (1969). Notice that in Liang and Wu (2008) and Varah (1982) the criterion function is given by

$$\sum_{i=1}^n \|\widetilde{x}'(t_i) - F(\widetilde{x}(t_i),\eta)\|^2,$$

where $\widetilde{x}$ and $\widetilde{x}'$ are appropriate estimators of x_θ and x_θ'. However, in order to obtain a $\sqrt{n}$-consistent estimator of θ, it is important to use an integral type criterion: the nonparametric estimators of x_θ and x_θ' have a slower convergence rate than $\sqrt{n}$ and this is counterbalanced by the integral criterion from (13). Indeed, stationarity at $\hat{\theta}_n$ leads to (37). The first factor at the left hand side of this equality converges to a constant nondegenerate matrix and the righthand side behaves like a linear combination of the observations with coefficients of order $1/n$ thanks to the integration; cf. Proposition 4 and its proof. In light of this the choice of the weight function w also appears to be important. Furthermore, the observations Y_{ij} from (5) indirectly carry information on the entire curves $x_{\theta j}(t), t \in [0,1]$, and not only on the points $x_{\theta j}(t_i)$. An integral type criterion allows one to exploit this fact in the second step of this smooth and match procedure.

Introduce the asymptotic criterion

$$M_w(\eta) = \int_0^1 \|F(x_\theta(t),\theta) - F(x_\theta(t),\eta)\|^2 w(t)dt$$

corresponding to $M_{n,w}$. Observe that by Condition 4 it is bounded. Using Corollary 1 as a building block, one can show that the SME $\hat{\theta}_n$ is consistent. To this end we will need the following condition on the weight function w.

Condition 6. *The weight function w is a nonnegative function that is continuously differentiable, is supported on the interval $(\delta, 1-\delta)$ for some fixed number δ, such* *that* $0 < \delta < 1/2$, *and is such that the Lebesgue measure of the set* $\{t : w(t) > 0\}$ *is positive.*

The fact that w vanishes at the endpoints of the interval $[\delta, 1-\delta]$ and beyond, is needed to obtain a $\sqrt{n}$-consistent estimator of θ. In particular, together with differentiability of w it is used in order to establish (40). The condition that w is supported on $(\delta, 1-\delta)$ takes care of the boundary bias effects characteristic of the conventional kernel-type estimators, see e.g. Gasser and Müller (1984) for more information on this. Boundary effects in kernel estimation are usually remedied by using special boundary kernels, see e.g. van Es (1991), Gasser et al. (1985), Messer and Goldstein (1993). Using such a kernel, it can be expected that in our case as well the boundary effects will be eliminated and one may relax the requirement $0 < \delta < 1/2$ from Condition 6 to $\delta = 0$, i.e. to allowing w to be supported on $(0, 1)$. The condition that the weight function w is positive on a set with positive Lebesgue measure, is important for (14) to hold and in fact $w(t) = 0$ a.e. would be a strange choice.

The following proposition is valid.

Proposition 2. *Suppose* $b \to 0$ *and* $nb^3/\log n \to \infty$. *Under Conditions 1–6 and the additional identifiability condition*

$$\forall \varepsilon > 0, \inf_{\|\eta-\theta\|\geq\varepsilon} M_w(\eta) > M_w(\theta), \tag{14}$$

we have $\hat{\theta}_n \overset{\mathrm{P}}{\to} \theta$.

The proposition is proved via a reasoning standard in the theory of M-estimation: we show that $M_{n,w}$ converges to M_w and that the convergence is strong enough to imply the convergence of a minimiser $\hat{\theta}_n$ of $M_{n,w}$ to a minimiser θ of M_w, cf. Section 5.2 of van der Vaart (1998). A necessary condition for (14) to hold is that $x_\theta(\cdot) \neq x_{\theta'}(\cdot)$ for $\theta \neq \theta'$. The latter is a minimal assumption for the statistical identifiability of the parameter θ. The identifiability condition (14) is common in the theory of M-estimation, see Theorem 5.7 of van der Vaart (1998). It means that θ is a point of minimum of $M_w(\eta)$ and that it is a *well-separated* point of minimum. The most trivial example with this condition satisfied is when $d = p = 1$ and $x'(t) = \theta x(t)$ hold with initial condition $x(0) = \xi$, where $\xi \neq 0$. In fact, in this case

$$M_w(\eta) = (\theta-\eta)^2\xi^2 \int_\delta^{1-\delta} e^{2\theta t} w(t) dt,$$

and this is zero for $\eta = \theta$ and is strictly positive for $\eta \neq \theta$, whence (14) follows. More generally, since Θ is compact and M_w is continuous, uniqueness of a minimiser of M_w will imply (14), cf. Exercise 27 on p. 84 of van der Vaart (1998).

In practice (14) might be difficult to check globally and one might prefer to concentrate on a simpler local condition: if the first order condition $[dM_w(\eta)/d\eta]_{\eta=\theta} = 0$ holds and if the Hessian matrix $H(\eta) = (\partial^2 M_w(\eta)/\partial\eta_i\partial\eta_j)_{i,j}$ of M_w is strictly positive definite at θ, then (14) will be satisfied for $\eta \in \Theta$ restricted to some neighbourhood of θ, because M_w will have a local minimum at such θ and a neighbourhood around it can be taken to be compact with small enough diameter, so that (14) holds for η restricted to this neighbourhood. The conclusion of the theorem will then hold for the parameter set restricted to this neighbourhood of θ.

In a statement analogous to Proposition 2, Brunel (2008) requires that the solutions of (4) belong to a compact set $\mathcal{K}$ for all θ and t and that F from (1) is Lipschitz in its first argument x for x restricted to this compact $\mathcal{K}$ uniformly in $\theta \in \Theta$. It is also assumed that the nonparametric estimators $\hat{x}_n(t)$ belong a.s. to $\mathcal{K}$ for all n and t. However, the latter typically will not hold for linear smoothers, see Definition 1.7 in Tsybakov (2009), which constitute the most popular choice of nonparametric regression estimators in practice. For instance, local polynomial estimators, see Section 1.6 in Tsybakov (2009), projection estimators, see Section 1.7 in Tsybakov (2009), or the Gasser-Müller estimator, see Gasser and Müller (1984), are all examples of linear smoothers. Hence we prefer to avoid this condition altogether, although this somewhat complicates the proof.

Under the conditions in this section it turns out that the estimator $\hat{\theta}_n$ is not merely a consistent estimator, but a $\sqrt{n}$-consistent estimator of θ, in the sense of (18) below. This result follows in essence from the fact that up to a higher order term the difference $\hat{\theta}_n - \theta$ can be represented as the difference of the images of $\hat{x}$ and x_θ under a certain linear mapping, cf. Proposition 3. It is known that even though nonparametric curve estimators cannot usually attain the $\sqrt{n}$ convergence rate, see e.g. Chapters 1 and 2 of Tsybakov (2009), extra smoothness often coming from the structure of linear functionals allows one to construct in many cases $\sqrt{n}$-consistent estimators of these functionals via plugging in nonparametric estimators, see e.g. Bickel and Ritov (2003) and Goldstein and Messer (1992) for more information. The variance of such plug-in estimators can often be proven to be of order n^{-1}, while the squared bias can be made of order n^{-1} by undersmoothing, i.e. selecting the smoothing parameter smaller than what is an optimal choice in nonparametric curve estimation when the object of interest is a curve itself, cf. Goldstein and Messer (1992). Precisely this happens in our case as well: if the mean integrated squared error is used as a performance criterion of a nonparametric estimator, then under our conditions the optimal bandwidth for estimation of x_θ is of order $n^{-1/(2\alpha+1)}$, whereas the optimal bandwidth for estimation of θ is in fact smaller, see Theorem 1 below. Note that undersmoothing is a different approach than the one in Bickel and Ritov (2003), where it is assumed that nonparametric estimators attain the minimax rate of convergence and the $\sqrt{n}$-rate for estimation of a functional in concrete examples, if possible, is achieved by different means exploiting extra smoothness coming from the structure of a functional, see e.g. the first example in Section 2 there. In many cases it can be proved that such plug-in type estimators are efficient, see Bickel and Ritov (2003). Notice, however, that in our case this will not imply that $\hat{\theta}_n$ is efficient.

First we will provide an asymptotic representation for the difference $\hat{\theta}_n - \theta$.

Proposition 3. *Let θ be an interior point of Θ. Suppose that the conditions of Proposition 2 hold and let the matrix J_θ defined by*

$$J_\theta = \int_\delta^{1-\delta} (F'_\theta(x_\theta(t),\theta))^T F'_\theta(x_\theta(t),\theta) w(t) dt \tag{15}$$

be nonsingular. Fix $\alpha \geq 3$. If $b \asymp n^{-\gamma}$ holds for $1/(4\alpha-4) < \gamma < 1/6$, then

$$\hat{\theta}_n - \theta = O_P\left(J_\theta^{-1}(\Gamma(\hat{x}) - \Gamma(x_\theta))\right) + o_P(n^{-1/2}) \tag{16}$$

is valid with the mapping Γ given by

(17)

$$\Gamma(z) = \int_\delta^{1-\delta} \left\{ -(F'_\theta(x_\theta(t),\theta))^T F'_x(x_\theta(t),\theta) w(t) - \frac{d}{dt}[(F'_\theta(x_\theta(t),\theta))^T w(t)] \right\} z(t) dt.$$

With the above result in mind, in order to complete the study of the asymptotics of $\hat{\theta}_n$, it remains to study the mapping Γ. Clearly, it suffices to study the asymptotic behaviour of

$$\Delta(\hat{\mu}_n) - \Delta(\mu) = \int_{\mathbb{R}} v(t)k(t)\hat{\mu}_n(t)dt - \int_{\mathbb{R}} v(t)k(t)\mu(t)dt,$$

where v is a known function that satisfies appropriate assumptions, while k stands either for w or its derivative w'. The next proposition deals with the asymptotics of $\Delta(\hat{\mu}_n) - \Delta(\mu)$.

Proposition 4. *Under Conditions 5 and 6 and for any continuous function v it holds in the regression model* (7) *that*

$$\Delta(\hat{\mu}_n) - \Delta(\mu) = O_P(n^{-1/2}),$$

provided μ is $\alpha \geq 3$ times differentiable and the bandwidth b is chosen such that $b \asymp n^{-\gamma}$ holds for $1/(2\alpha) \leq \gamma \leq 1/4$.

Our main result is a simple consequence of Propositions 3 and 4.

Theorem 1. *Let θ be an interior point of Θ. Assume that Conditions 1–6 together with* (14) *hold and that* (15) *is nonsingular. Fix $\alpha \geq 4$. If the bandwidth b is such that $b \asymp n^{-\gamma}$ holds for $1/(2\alpha) < \gamma < 1/6$, then*

$$\sqrt{n}(\hat{\theta}_n - \theta) = O_P(1) \tag{18}$$

is valid.

Thus any bandwidth sequences satisfying the conditions in Theorem 1 are optimal, in the sense that they lead to estimators of θ with similar asymptotic behaviour. In particular, each of such bandwidth sequences ensures a $\sqrt{n}$ convergence rate of $\hat{\theta}_n$. Consequently, dependence of the asymptotic properties of the estimator $\hat{\theta}_n$ on the bandwidth is less critical than it typically is in nonparametric curve estimation. Notice that the condition $\alpha \geq 4$ in Theorem 1 is needed in order to make the conditions in Propositions 3 and 4 compatible.

4. Discussion

The main result of the paper, Theorem 1, is that under certain conditions for systems of ordinary differential equations parameter estimation at the $\sqrt{n}$ rate is possible *without* employing numerical integration. Although we have shown this in the case when in the first step of the two-step procedure a particular kernel-type estimator is used, it may be expected that a similar result holds for other nonparametric estimators. For instance, the arguments for the Nadaraya-Watson estimator seem to be similar, with extra technicalities arising e.g. from the fact that it is a ratio of two functions. Furthermore, from formula (40) it can be seen that the proof of Proposition 3 requires that the derivative of an estimator of x_θ be used as an estimator of x'_θ. Not all popular nonparametric estimators of the derivatives of a regression function are of this type. In practice for small or moderate sample sizes it might be advantageous to use more sophisticated nonparametric estimators than the Priestley-Chao estimator, but asymptotically this does not make a difference.

Once a $\sqrt{n}$-consistent estimator $\hat{\theta}_n$ of θ is available, one might ask for more, namely if one can construct an estimator that is asymptotically equivalent to the ordinary least squares estimator (2) or that is semiparametrically efficient. It is expected that this can be achieved without repeated numerical integration of (1) by using $\hat{\theta}_n$ as a starting point and performing a one-step Newton-Raphson type procedure; see e.g. Section 7.8 of Bickel et al. (1998) or Chapter 25 of van der Vaart (1998). We intend to address this issue of efficient and ordinary least squares estimation in a separate publication.

Doubtless, the main challenge in implementing the smooth and match estimation procedure lies in selecting the smoothing parameter b. This is true for any two-step parameter estimation procedure for ordinary differential equations, e.g. the one based on the regression splines as in Brunel (2008) or the local polynomial estimator as in Liang and Wu (2008), and not only for our specific estimator. Observations that we supply below apply in principle to any two-step estimator and not only to the specific kernel-type one considered in the present work. Hence they are of general interest.

Some attention has been paid in the literature to the selection of the smoothing parameter in the context of parameter estimation for ordinary differential equations. The considered options range from subjective choices and smoothing by hand to more advanced possibilities. Perhaps the simplest solution would be to assume that the targets of the estimation procedure are $x_{\theta j}$, $j = 1, \ldots, d$, and to select b (a different one for every component $x_{\theta j}$) via a cross-validation procedure, see e.g. Section 5.3 in Wasserman (2006) for a description of cross-validation techniques in the context of nonparametric regression. This should produce reasonable results, at least for relatively large sample sizes, cf. simulation examples considered in Brunel (2008). However, it is clear from Theorem 1 and its proof that despite its simplicity, such a choice of b will be suboptimal. Another practical approach to bandwidth selection is computation of $\hat{\theta}_n = \hat{\theta}_n(b)$ for a range of values of the bandwidth b on some discrete grid B and then choosing

$$\hat{b} = \operatorname{argmin}_{b \in B} \sum_{i=1}^{n} \sum_{j=1}^{d} (Y_{ij} - x_{\hat{\theta}_n(b)j}(t_i))^2.$$

This seems a reasonable choice, although the asymptotics of $\hat{\theta}_n(\hat{b})$ are unclear. One other possibility for practical bandwidth selection is nothing else but a variation on the plug-in bandwidth selection method as described e.g. in Jones et al. (1996): one can see from the proof in Section 5 that the terms that depend on the bandwidth b are lower order terms in the expansion of $\hat{\theta}_n - \theta$. One can then minimise with respect to b a bound on these lower order terms. A minimiser, say b^*, will depend on the unknown true parameter θ, also via x_θ and x_θ', as well as on the error variances $\sigma_1^2, \ldots, \sigma_d^2$. However, θ, x_θ, and x_θ' can be re-estimated via $\hat{\theta}_n, \hat{x}$, and $\hat{x}'$ using a different, pilot bandwidth $\tilde{b}$. Of course, instead of $\hat{x}$ and $\hat{x}'$ the use of any other nonparametric estimators of a regression function and its derivative, e.g. local polynomial estimators, see Section 1.6 of Tsybakov (2009), or the Gasser-Müller estimator, see Gasser and Müller (1984), is also a valid option. Error term variances can be estimated via one of the methods described in Hall and Marron (1990) or Section 5.6 of Wasserman (2006). Once the pilot estimators of θ, x_θ, and x_θ' together with estimators of $\sigma_1^2, \ldots, \sigma_d^2$ are available, these can be plugged back into b^* and in this way one obtains a bandwidth $\hat{b}$ that estimates the optimal bandwidth b^*. The final step would be computation of $\hat{\theta}_n$ with a new bandwidth $\hat{b}$. Unfortunately, this method leads to extremely cumbersome expressions and furthermore, since we are minimising an upper bound on numerous remainder terms, it will probably tend to oversmooth, i.e. produce a bandwidth b larger than required. Moreover, the plug-in approach in general is subject to some controversy having both supporters and critics, see e.g. Loader (1999) and references therein. An alternative to the plug-in approach might be an approach based on one of the resampling methods: cross-validation, jackknife, or bootstrap. Computationally these resampling methods will be quite intensive. Theoretical analysis of the properties of such bandwidth selectors is a rather nontrivial task. Also a thorough simulation study is needed before the practical value of different bandwidth selection methods can be assessed. We do not address these issues here.

The next observation of this section concerns numerical computation of our SME. The kernel-type nonparametric regression estimates of $x_{\theta j}$, $j = 1, \ldots, d$, can be quickly evaluated on any regular grid of points $0 \leq s_1 \leq \ldots \leq s_m$, e.g. via techniques using the Fast Fourier Transform (FFT) similar to those described in Appendix D of Wand and Jones (1995). See also Fan and Marron (1994). Furthermore, in the match step of the two-step estimation procedure the criterion function $M_{n,w}$ can be approximated by a finite sum by discretising the integral in its definition. If F is linear in $\theta_1, \ldots, \theta_p$ and is univariate, then as already observed in Varah (1982), see pp. 29 and 31, cf. p. 1262 in Brunel (2008) and p. 1573 in Liang and Wu (2008), this will lead to a weighted linear least squares problem, which can be solved in a routine fashion without using e.g. random search methods. This is a great simplification in comparison to the ordinary least squares estimator, which moreover will still tend to get trapped in local minima of the least squares criterion function despite the fact that F is linear in its parameters.

We conclude this section with two simple problems illustrating parameter estimation for systems of ordinary differential equations via the smooth and match method studied in the present paper. Our first example deals with the Lotka-Volterra system that is a basic model in population dynamics. It describes evolution over time of the populations of two species, predators and their preys. In mathematical terms the Lotka-Volterra model is described by a system consisting of two ordinary differential equations and depending on the parameter $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)^T$,

$$\begin{cases} x_1'(t) = \theta_1 x_1(t) - \theta_2 x_1(t) x_2(t), \\ x_2'(t) = -\theta_3 x_2(t) + \theta_4 x_1(t) x_2(t). \end{cases} \tag{19}$$

Here x_1 represents the prey population and x_2 the predator population. For additional information on the Lotka-Volterra system see e.g. Section 6.2 in Edelstein-Keshet (2005). We took $\theta_k = 0.5, k = 1, \ldots, 4$ and the initial condition $(x_1(0), x_2(0)) = (1, 0.5)$. The solution to (19) corresponding to these parameter values is plotted in Figure 1 with a thin line. The left panel represents $x_{\theta 1}$, the right panel $x_{\theta 2}$. The solution components $x_{\theta 1}$ and $x_{\theta 2}$ are of oscillatory nature and are out of phase of each other. Next we simulated a small data set of size $n = 50$ of observations on the solution x_θ of (19) over the time interval $[0, 25]$ by taking an equidistant grid of time points $t_i = 0.5i$ for $i = 1, \ldots, 50$ and setting

$$Y_{ij} = x_{\theta j}(t_i) + \epsilon_{ij}, \quad i = 1, \ldots, 50, j = 1, 2, \tag{20}$$

where the i.i.d. measurement errors ϵ_{ij} were generated from the normal distribution $N(0, \sigma^2)$ with mean zero and variance $\sigma^2 = 0.01$. These observations Y_{ij} are represented by crosses in Figure 1.

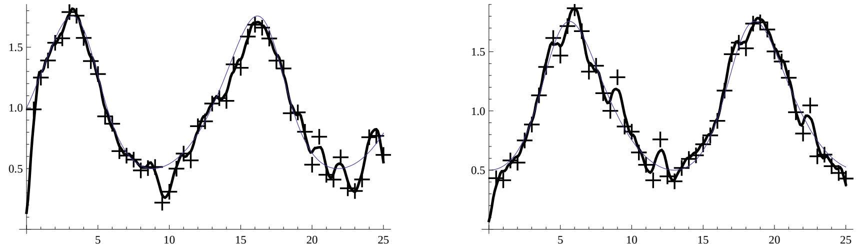

FIGURE 1. Solution of the Lotka-Volterra system (19) (thin line) with parameter values $\theta_k = 0.5, k = 1, \ldots, 4$, and initial condition $(x_1(0), x_2(0)) = (1, 0.5)$, observations Y_{ij} given by (20) with $\epsilon_{ij} \sim N(0, 0.01)$ (crosses) and the estimates $\hat{x}_j$ computed with kernel (21), weight function (22) and bandwidth $b = 1.2$ (solid line). The left panel corresponds to $x_{\theta 1}$, the right to $x_{\theta 2}$.

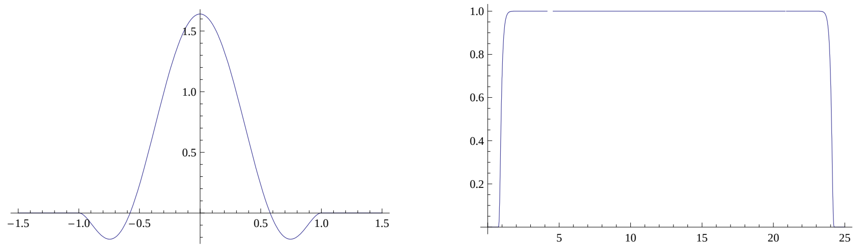

FIGURE 2. Kernel K from (21) (left panel) and weight function w from (22) (right panel).

The three required ingredients for the construction of an estimator $\hat{\theta}_n$ are the kernel K, the weight function w, and the bandwidth b. A general recipe for construction of kernels of an arbitrary order α is given in Section 1.2.2 of Tsybakov (2009) and is based on the use of polynomials that are orthonormal in $L_2(-1, 1)$ with weights. In particular, we used the ultraspherical or Gegenbauer polynomials with weight function $v(t) = (1-t^2)^2 1_{[|t|\leq 1]}$ and constructed the fourth order kernel with them. Notice that our definition of the kernel of order α in Condition 5 is slightly different from the one in Definition 1.3 of Tsybakov (2009), cf. also the remark on p. 6 there. For ultraspherical polynomials see Section 4.7 in Szegö (1975). Our fourth order kernel took the form

$$K(t) = \left(\frac{105}{64} - \frac{315}{64} t^2 \right) (1 - t^2)^2 1_{[|t|\leq 1]}. \tag{21}$$

Notice that K is a symmetric function. The kernel K is plotted in Figure 2 in the left panel. An alternative here is to use the Gaussian-based kernels as in Wand and Schucany (1990), although they do not have a compact support. As far as the weight function w is concerned, any nonnegative function that is equal to zero close to the end points of the interval $[0, 25]$, is equal to one on the greater part of the interval $[0, 25]$ and is smooth, could have been used. We opted to simply

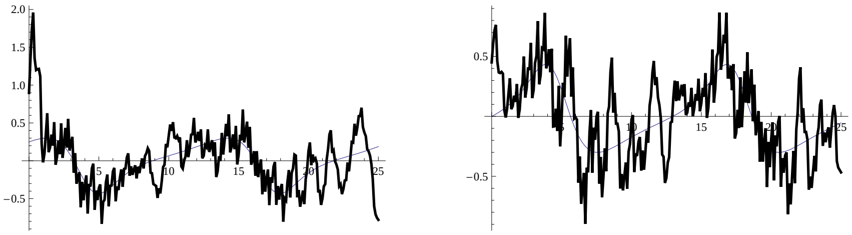

FIGURE 3. Derivatives of the solution components $x_{\theta j}$ of the Lotka-Volterra system (19) (thin line) with parameter values $\theta_k = 0.5, k = 1, \ldots, 4$, and initial condition $(x_1(0), x_2(0)) = (1, 0.5)$, together with derivative estimates $\hat{x}_j'$ (solid line) computed with kernel (21), weight function (22), and bandwidth $b = 1.2$ using observations Y_{ij} from (20). The left panel corresponds to $\hat{x}_1'$, the right panel to $\hat{x}_2'$.

rescale and shift the function

$$\lambda_{c,\beta}(t) = \begin{cases} 1, & \text{if } |t| \leq c, \\ \exp[-\beta \exp[-\beta/(|t|-c)^2]/(|t|-1)^2], & \text{if } c < |t| < 1, \\ 0, & \text{if } |t| \geq 1, \end{cases}$$

that arose in a different context in McMurry and Politis (2004), see formula (3) on p. 552 there, so that it could have the required properties in our context. We took the constants c and β to be equal to 0.7 and 0.5, respectively, and then set

$$w(t) = \lambda_{c,\beta}\left(1.05\frac{(t-12.5)}{12.5}\right). \tag{22}$$

The function w is plotted in the right panel of Figure 2. Finally, since in the present work construction of the bandwidth selector is not our primary goal, we simply selected b by hand and set it to 1.2.

The smooth and match estimation procedure was implemented in *Mathematica* 6.0, see Wolfram Research, Inc. Mathematica, Version 6.0 (2007). We first evaluated the kernel estimates of the regression functions $x_{\theta 1}$ and $x_{\theta 2}$ at the equidistant grid of points $s_k = 0.1k$ with $k = 0, \ldots, 249$. With this number of grid points and the sample size $n = 50$ there was no need to use binning to compute the estimates and moreover, binning would have probably resulted in a slower procedure, cf. Figure 3b in Fan and Marron (1994); so we did not employ it. However, the fact that many of the kernel evaluations $K((s_k - t_i)/b)$ are actually the same, cf. Fan and Marron (1994), was taken into account and led to savings in computation time above the naive implementation of the Priestley-Chao estimator that would directly compute $K((t - t_i)/b)$. The estimates $\hat{x}_1$ and $\hat{x}_2$ are plotted in Figure 1 with a solid line, while the estimates $\hat{x}_1'$ and $\hat{x}_2'$ are plotted in Figure 3. Notice that the estimates $\hat{x}_1'$ and $\hat{x}_2'$ are severely undersmoothed. We next approximated the criterion function $M_{n,w}$ by a Riemann sum

$$\sum_{k=0}^{249} (\hat{x}_1'(0.1k) - \eta_1 \hat{x}_1(0.1k) + \eta_2 \hat{x}_1(0.1k)\hat{x}_2(0.1k))^2 w(0.1k) 0.1$$

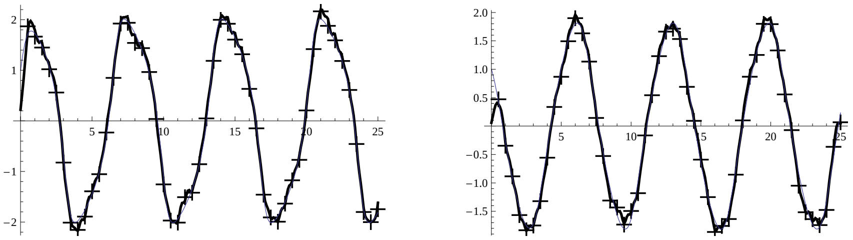

Figure 4. Solution of the Van der Pol system (23) (thin line) with parameter value $\theta = 0.8$ and initial condition $(x_1(0), x_2(0)) = (1, 1)$, observations Y_{ij} given by (24) with $\epsilon_{ij} \sim N(0, 0.01)$ (crosses) and the estimates $\hat{x}_j$ computed with kernel (21), weight function (22), and bandwidth $b = 1$ (solid line). The left panel corresponds to $x_{\theta 1}$ and the right to $x_{\theta 2}$.

$$+ \sum_{k=0}^{249} (\hat{x}_2'(0.1k) + \eta_3 \hat{x}_2(0.1k) - \eta_4 \hat{x}_1(0.1k)\hat{x}_2(0.1k))^2 w(0.1k) 0.1.$$

Note that when performing minimisation, the factor 0.1 can be omitted from both terms in the above display. The minimisation procedure resulted in the estimate

$$\hat{\theta}_n = (0.52, 0.50, 0.50, 0.51)^T.$$

With our implementation, the total time needed for computation of the estimate of θ (including time needed for kernel and weight function evaluations, but excluding time needed for loading observations) was about 0.5 seconds on a notebook with Intel(R) Pentium(R) Dual CPU T3200 @ 2.00 GHz processor and 4.00 GB RAM. The parameter estimates appear to be sufficiently accurate in this particular case.

Our second example deals with the Van der Pol oscillator that describes an electric circuit containing a nonlinear element, see p. 333, Problem 12 on p. 365, and the references on p. 373 in Edelstein-Keshet (2005). The corresponding system of ordinary differential equations takes the form

$$\begin{cases} x_1'(t) = \theta^{-1} \left(x_1(t) - \frac{1}{3}(x_1(t))^3 + x_2(t) \right), \\ x_2'(t) = -\theta x_1(t). \end{cases} \tag{23}$$

We took $\theta = 0.8$ and the initial condition $(x_1(0), x_2(0)) = (1, 1)$. The solution to (23) is of oscillatory nature and the components $x_{\theta 1}$ and $x_{\theta 2}$ are out of phase of each other. The solution is plotted in Figure 4 with a thin line. We then simulated a data set of size $n = 50$ of observations on the solution x_θ of (23) over the time interval $[0, 25]$ at an equidistant grid of time points $t_i = 0.5i, i = 1, \dots, 50$, by setting

$$Y_{ij} = x_{\theta j}(t_i) + \epsilon_{ij}, \quad i = 1, \dots, 50, j = 1, 2, \tag{24}$$

where the i.i.d. measurement errors ϵ_{ij} were generated from the normal distribution $N(0, \sigma^2)$ with mean zero and variance $\sigma^2 = 0.01$. These observations Y_{ij} are plotted with crosses in Figure 4. When computing the estimate $\hat{\theta}_n$, we used the same kernel and the same weight function as in the previous example, while the bandwidth was set to $b = 1$. The estimates of the solution components $x_{\theta 1}$ and $x_{\theta 2}$ are depicted by a solid line in Figure 4, while the derivatives $x_{\theta 1}'$ and $x_{\theta 2}'$ together with their

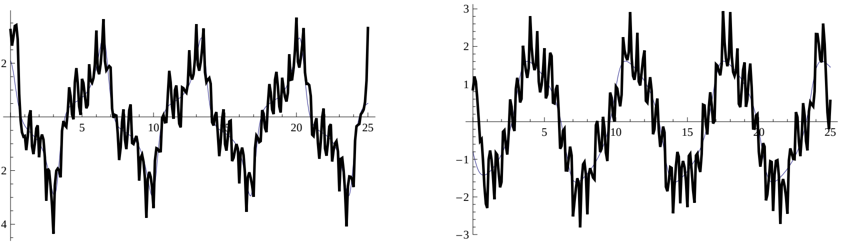

FIGURE 5. Derivatives of the solution components $x_{\theta j}$ of the Van der Pol system (23) (thin line) with parameter value $\theta = 0.8$ and initial condition $(x_1(0), x_2(0)) = (1, 1)$, together with derivative estimates $\hat{x}_j'$ (solid line) computed with kernel (21), weight function (22), and bandwidth $b = 1$ using observations Y_{ij} from (24). The left panel corresponds to $\hat{x}_1'$ and the right panel to $\hat{x}_2'$.

estimates are given in Figure 5. The estimation procedure resulted in an estimate $\hat{\theta}_n = 0.83$ and the computation time was about 0.4 seconds.

We intend to perform a more practically oriented study exploring some of the ideas mentioned in this section in a separate publication.

5. PROOFS

We will use the symbol $\lesssim$, meaning less or equal up to a universal constant independent of index n. The symbol $\asymp$ will denote the fact that two sequences of real numbers are asymptotically of the same order.

Proof of Proposition 1. We first prove (9). For any positive ε by Chebyshev's inequality we have

$$\begin{aligned}P\left(\sup_{t\in[\delta,1-\delta]}|\hat{\mu}_n(t)-\mu(t)|>\varepsilon\right) &\leq \frac{2}{\varepsilon^2}\Bigg\{\sup_{t\in[\delta,1-\delta]}|\mathbb{E}\left[\hat{\mu}_n(t)\right]-\mu(t)|^2\\ &\qquad+\mathbb{E}\left[\sup_{t\in[\delta,1-\delta]}|\hat{\mu}_n(t)-\mathbb{E}\left[\hat{\mu}_n(t)\right]|^2\right]\Bigg\}\\ &=\frac{2}{\varepsilon^2}(T_1+T_2).\end{aligned}\tag{25}$$

By (45) we can write

$$\mathbb{E}\left[\hat{\mu}_n(t)\right]-\mu(t)=\int_0^1\mu(s)\frac{1}{b}K\left(\frac{t-s}{b}\right)ds-\mu(t)+O\left(\frac{1}{nb^2}\right).$$

For all n large enough, we have $b \leq \delta$, because $b \to 0$. Then for all such n, if $t \in [\delta, 1-\delta]$, a standard argument (cf. p. 6 in Tsybakov (2009)), namely Taylor's formula up to order α applied to μ and the moment conditions on the kernel K formulated in Condition 5, yields

$$\sup_{t\in[\delta,1-\delta]}|\mathbb{E}\left[\hat{\mu}_n(t)\right]-\mu(t)|\leq b^\alpha\frac{\|\mu^{(\alpha)}\|_\infty}{\alpha!}\int_{-1}^1|u^\alpha K(u)|du+O\left(\frac{1}{nb^2}\right).\tag{26}$$

Next we turn to T_2. With argumentation similar to that in the proof of Theorem 1.8 of Tsybakov (2009) and setting

$$S_i(t) = \frac{t_i - t_{i-1}}{b} K\left(\frac{t-t_i}{b}\right), \quad N = n^2, \quad s_j = \frac{j}{N},$$

for $j = 1, \ldots, N$, we have

$$\begin{aligned} A &= \sup_{t\in[\delta,1-\delta]} |\hat{\mu}_n(t) - \mathbb{E}[\hat{\mu}_n(t)]| \\ &= \sup_{t\in[\delta,1-\delta]} \left|\sum_{i=1}^n S_i(t)\epsilon_i\right| \\ &\leq \max_{1\leq j\leq N} \left|\sum_{i=1}^n S_i(s_j)\epsilon_i\right| + \sup_{t,t':|t-t'|\leq N^{-1}} \left|\sum_{i=1}^n (S_i(t) - S_i(t'))\epsilon_i\right|. \end{aligned}$$

By the mean value theorem and Condition 1 the inequality

$$|S_i(t) - S_i(t')| \lesssim \|K'\|_\infty \frac{1}{nb^2}|t-t'|$$

holds for any $t, t' \in \mathbb{R}$, where $\|K'\|_\infty$ is finite. Hence by the c_2-inequality

$$\begin{aligned} A^2 &\leq \left(\max_{1\leq j\leq N} \left|\sum_{i=1}^n \epsilon_i S_i(s_j)\right| + \sup_{t,t':|t-t'|\leq N^{-1}} \left|\sum_{i=1}^n (S_i(t) - S_i(t'))\epsilon_i\right|\right)^2 \\ &\lesssim \max_{1\leq j\leq N} |Z_j|^2 + \frac{\|K'\|_\infty^2}{n^2b^4N^2}\left(\sum_{i=1}^n |\epsilon_i|\right)^2, \end{aligned} \tag{27}$$

where $Z_j = \sum_{i=1}^n S_i(s_j)\epsilon_i$. Notice that

$$\frac{1}{n^2b^4N^2}\mathbb{E}\left[\left(\sum_{i=1}^n |\epsilon_i|\right)^2\right] \leq \frac{\mathbb{E}[\epsilon_1^2]}{N^2b^4} = \frac{\sigma^2}{n^4b^4} = o\left(\frac{1}{nb}\right). \tag{28}$$

Moreover, we have

$$\begin{aligned} \mathbb{E}[Z_j^2] &= \sum_{i=1}^n \sigma^2(t_i - t_{i-1})^2\left(\frac{1}{b}K\left(\frac{t_i - s_j}{b}\right)\right)^2 \\ &\lesssim \frac{\sigma^2\|K\|_\infty^2}{n^2b^2}\sum_{i=1}^n 1_{[|t_i-s_j|\leq b]} \\ &\leq \frac{1}{nb}c_1\sigma^2\|K\|_\infty^2 \max\left(2, \frac{1}{nb}\right), \end{aligned}$$

where the last inequality follows from Condition 1. Since the Z_j's, being a linear combination of independent Gaussian random variables, are themselves Gaussian, Corollary 1.3 of Tsybakov (2009) and the fact that $N = n^2$ then entail

$$\mathbb{E}\left[\max_{1\leq j\leq N} |Z_j|^2\right] = O\left(\frac{\log N}{nb}\right) = O\left(\frac{\log n}{nb}\right). \tag{29}$$

Combining (27), (28) and (29), we obtain

$$\mathbb{E}[A^2] = O\left(\frac{\log n}{nb}\right). \tag{30}$$

Taking

$$\varepsilon = M\left(b^\alpha + \frac{1}{nb^2} + \sqrt{\frac{\log n}{nb}}\right)$$

with an appropriate constant M yields (9) by (25), (26), and (30).

As far as the proof of (10) is concerned, it is very much similar to the proof of (9) and is therefore omitted. This completes the proof of the proposition. □

Proof of Proposition 2. From the definition of $M_{n,w}(\eta)$ and $M_w(\eta)$, the elementary inequality

$$|\|a_1\|^2 - \|a_2\|^2| \le \|a_1 - a_2\|(\|a_1\| + \|a_2\|)$$

and the Cauchy-Schwarz inequality we have

$$\begin{aligned} &|M_{n,w}(\eta) - M_w(\eta)| \\ &\le \left\{\int_0^1 \|\hat{x}'(t) - F(x_\theta(t),\theta) + F(x_\theta(t),\eta) - F(\hat{x}(t),\eta)\|^2 w(t)dt\right\}^{1/2} \\ &\times \left\{\sqrt{\int_0^1 \|\hat{x}'(t) - F(\hat{x}(t),\eta)\|^2 w(t)dt} + \sqrt{\int_0^1 \|F(x_\theta(t),\theta) - F(x_\theta(t),\eta)\|^2 w(t)dt}\right\} \\ &= \sqrt{T_1}(\sqrt{T_2} + \sqrt{T_3}). \end{aligned} \tag{31}$$

For T_1 we have that

$$\begin{aligned} T_1 \le 2\int_\delta^{1-\delta} \|\hat{x}'(t) - F(x_\theta(t),\theta)\|^2 w(t)dt \\ + 2\int_\delta^{1-\delta} \|F(x_\theta(t),\eta) - F(\hat{x}(t),\eta)\|^2 w(t)dt. \end{aligned} \tag{32}$$

By (12) it holds that

$$\begin{aligned} \sup_{\eta\in\Theta} \int_\delta^{1-\delta} \|\hat{x}'(t) - F(x_\theta(t),\theta)\|^2 w(t)dt \\ &= \int_\delta^{1-\delta} \|\hat{x}'(t) - x'_\theta(t)\|^2 w(t)dt \\ &\le \sum_{i=1}^d \sup_{t\in[\delta,1-\delta]} |\hat{x}'_i(t) - x'_{i,\theta}(t)|^2 \int_\delta^{1-\delta} w(t)dt \\ &\overset{\mathrm{P}}{\to} 0. \end{aligned} \tag{33}$$

Moreover, by Lemma 3 from Appendix A we obtain that

$$\sup_{\eta\in\Theta} \int_\delta^{1-\delta} \|F(\hat{x}(t),\eta) - F(x_\theta(t),\eta)\|^2 w(t)dt \overset{\mathrm{P}}{\to} 0. \tag{34}$$

Furthermore, $T_3 = O_P(1)$ as $n \to \infty$, because

$$\sup_{\eta\in\Theta} \int_\delta^{1-\delta} \|F(x_\theta(t),\theta) - F(x_\theta(t),\eta)\|^2 w(t)dt < \infty \tag{35}$$

by compactness of Θ and Condition 4, and $T_2 = O_P(1)$, because

$$\sup_{\eta\in\Theta}\int_\delta^{1-\delta}\|\hat{x}'(t)-F(\hat{x}(t),\eta)\|^2 w(t)dt = O_P(1) \tag{36}$$

holds by the inequality

$$\begin{aligned}\int_\delta^{1-\delta}&\|\hat{x}'(t)-F(\hat{x}(t),\eta)\|^2 w(t)dt\\ &\lesssim \int_\delta^{1-\delta}\|\hat{x}'(t)-x_\theta'(t)\|^2 w(t)dt + \int_\delta^{1-\delta}\|x_\theta'(t)-F(x_\theta(t),\eta)\|^2 w(t)dt\\ &\qquad + \int_\delta^{1-\delta}\|F(x_\theta(t),\eta)-F(\hat{x}(t),\eta)\|^2 w(t)dt,\end{aligned}$$

Corollary 1, compactness of Θ, Condition 4, and Lemma 3 from Appendix A. Combination of (31)–(36) implies that

$$\sup_{\eta\in\Theta}|M_{n,w}(\eta)-M_w(\eta)| \overset{\mathrm{P}}{\to} 0.$$

The statement of the proposition then follows from this fact, the identifiability condition (14), and Theorem 5.7 of van der Vaart (1998). □

Proof of Proposition 3. We interpret the derivative of a one-dimensional function of θ as a row p-vector of partial derivatives and we denote the $d\times p$-matrix of partial derivatives $\partial F_i(x,\theta)/\partial\theta_j$, $i=1,\dots,d$, $j=1,\dots,p$, by $F_\theta'(x,\theta)$.

We have

$$\frac{d}{d\theta}\|\hat{x}'(t)-F(\hat{x}(t),\theta)\|^2 = -2(\hat{x}'(t)-F(\hat{x}(t),\theta))^T F_\theta'(\hat{x}(t),\theta).$$

With this in mind and interchanging the order of integration and differentiation, we find that the derivative of $M_{n,w}$ from (13) with respect to θ is given by

$$-2\int_\delta^{1-\delta}(\hat{x}'(t)-F(\hat{x}(t),\theta))^T F_\theta'(\hat{x}(t),\theta)w(t)dt.$$

Since θ is an interior point of Θ, there exists an $\varepsilon>0$, such that the open ball of radius ε around θ is contained in Θ. Take

$$G_n = \{|\hat{\theta}_n-\theta|<\varepsilon/2\}$$

and notice that by consistency of $\hat{\theta}_n$ we have $P(G_n)\to 1$ as $n\to\infty$. If $\hat{\theta}_n$ is a point of minimum of $M_{n,w}$, then necessarily

$$1_{G_n}\int_\delta^{1-\delta}(\hat{x}'(t)-F(\hat{x}(t),\hat{\theta}_n))^T F_\theta'(\hat{x}(t),\hat{\theta}_n)w(t)dt = 0,$$

where 0 at the righthand side denotes now a row p-vector with all its entries equal to zero. The latter display can be rearranged as

$$\begin{aligned}1_{G_n}\int_\delta^{1-\delta}&(F_\theta'(\hat{x}(t),\hat{\theta}_n))^T\times\{(\hat{x}'(t)-x_\theta'(t))\\ &+(F(x_\theta(t),\theta)-F(\hat{x}(t),\theta))+(F(\hat{x}(t),\theta)-F(\hat{x}(t),\hat{\theta}_n))\}w(t)dt = 0,\end{aligned}$$

where now 0 on the righthand side denotes a column p-vector with its entries equal to zero. Note that we have

$$F(\hat{x}(t),\theta)-F(\hat{x}(t),\hat{\theta}_n)=\int_0^1 F'_\theta(\hat{x}(t),\hat{\theta}_n+\lambda(\theta-\hat{\theta}_n))d\lambda\,(\theta-\hat{\theta}_n).$$

Hence

$$\begin{aligned}1_{G_n}\int_\delta^{1-\delta}(F'_\theta(\hat{x}(t),\hat{\theta}_n))^T\int_0^1 F'_\theta(\hat{x}(t),\hat{\theta}_n+\lambda(\theta-\hat{\theta}_n))d\lambda\,w(t)dt\,(\hat{\theta}_n-\theta)&\\ =1_{G_n}\int_\delta^{1-\delta}(F'_\theta(\hat{x}(t),\hat{\theta}_n))^T(\hat{x}'(t)-x'_\theta(t))w(t)dt&\\ +1_{G_n}\int_\delta^{1-\delta}(F'_\theta(\hat{x}(t),\hat{\theta}_n))^T(F(x_\theta(t),\theta)-F(\hat{x}(t),\theta))w(t)dt&\end{aligned}\tag{37}$$

holds. By the fact that $\hat{x}$ converges in probability as a random element on $[\delta,1-\delta]$ to x_θ, see (11), consistency of $\hat{\theta}_n$, continuity of F'_θ, continuity of integration and the continuous mapping theorem, see Theorem 18.11 in van der Vaart (1998), we have

$$\begin{aligned}\int_\delta^{1-\delta}(F'_\theta(\hat{x}(t),\hat{\theta}_n))^T\int_0^1 F'_\theta(\hat{x}(t),\hat{\theta}_n+\lambda(\theta-\hat{\theta}_n))d\lambda\,w(t)dt&\\ \xrightarrow{\mathrm{P}}\int_\delta^{1-\delta}(F'_\theta(x_\theta(t),\theta))^T F'_\theta(x_\theta(t),\theta)w(t)dt=J_\theta,&\end{aligned}\tag{38}$$

where J_θ is nonsingular by assumption (15). Therefore, (37) shows that the asymptotic behaviour of $\hat{\theta}_n-\theta$ is given by

$$\begin{aligned}J_\theta^{-1}\Bigg(\int_\delta^{1-\delta}(F'_\theta(\hat{x}(t),\hat{\theta}_n))^T(\hat{x}'(t)-x'_\theta(t))w(t)dt&\\ +\int_\delta^{1-\delta}(F'_\theta(\hat{x}(t),\hat{\theta}_n))^T(F(x_\theta(t),\theta)-F(\hat{x}(t),\theta))w(t)dt\Bigg).&\end{aligned}\tag{39}$$

It thus remains to be shown that this expression in fact reduces to the righthand side of (16). First of all, notice that

$$\begin{aligned}&\int_\delta^{1-\delta}(F'_\theta(\hat{x}(t),\hat{\theta}_n))^T(\hat{x}'(t)-x'_\theta(t))w(t)dt\\ &\quad=\int_\delta^{1-\delta}(F'_\theta(x_\theta(t),\theta))^T(\hat{x}'(t)-x'_\theta(t))w(t)dt\\ &\quad\quad+\int_\delta^{1-\delta}(F'_\theta(\hat{x}(t),\hat{\theta}_n)-F'_\theta(x_\theta(t),\theta))^T(\hat{x}'(t)-x'_\theta(t))w(t)dt\\ &\quad=-\int_\delta^{1-\delta}\left(\frac{d}{dt}[F'_\theta(x_\theta(t),\theta)w(t)]\right)^T(\hat{x}(t)-x_\theta(t))dt\\ &\quad\quad+\int_\delta^{1-\delta}(F'_\theta(\hat{x}(t),\hat{\theta}_n)-F'_\theta(x_\theta(t),\theta))^T(\hat{x}'(t)-x'_\theta(t))w(t)dt,\end{aligned}\tag{40}$$

where the last equality follows by integration by parts and the fact that $w(\delta)=w(1-\delta)=0$. The first term at the righthand side of (40) appears also in the leading term $\Gamma(\hat{x}) - \Gamma(x_\theta)$ of (16). We will now show that the other term at the righthand side of (40) is negligible, i.e.

$$\int_\delta^{1-\delta} (F_\theta'(\hat{x}(t), \hat{\theta}_n) - F_\theta'(x_\theta(t), \theta))^T (\hat{x}'(t) - x_\theta'(t)) w(t) dt = o_P(n^{-1/2}).$$

By the Cauchy-Schwarz inequality

$$\begin{aligned}\left\| \int_\delta^{1-\delta} (F_\theta'(\hat{x}(t), \hat{\theta}_n) - F_\theta'(x_\theta(t), \theta))^T (\hat{x}'(t) - x_\theta'(t)) w(t) dt \right\| \\ \leq \left\{ \int_\delta^{1-\delta} \| F_\theta'(\hat{x}(t), \hat{\theta}_n) - F_\theta'(x_\theta(t), \theta) \|^2 w(t) dt \right\}^{1/2} \\ \times \left\{ \int_\delta^{1-\delta} \| \hat{x}'(t) - x_\theta'(t) \|^2 w(t) dt \right\}^{1/2},\end{aligned}$$

where $\| \cdot \|$ denotes the Frobenius or the Hilbert-Schmidt norm of a matrix (recall that it is submultiplicative). By (12) we have

$$\left\{ \int_\delta^{1-\delta} \| \hat{x}'(t) - x_\theta'(t) \|^2 w(t) dt \right\}^{1/2} = O_P(1) \left(b^{\alpha-1} + \frac{1}{nb^3} + \sqrt{\frac{\log n}{nb^3}} \right).$$

Furthermore,

$$\begin{aligned}\int_\delta^{1-\delta} & \| F_\theta'(\hat{x}(t), \hat{\theta}_n) - F_\theta'(x_\theta(t), \theta) \|^2 w(t) dt \\ & \leq 2 \int_\delta^{1-\delta} \| F_\theta'(\hat{x}(t), \hat{\theta}_n) - F_\theta'(x_\theta(t), \hat{\theta}_n) \|^2 w(t) dt \\ & + 2 \int_\delta^{1-\delta} \| F_\theta'(x_\theta(t), \hat{\theta}_n) - F_\theta'(x_\theta(t), \theta) \|^2 w(t) dt \\ & = 2T_1 + 2T_2. \end{aligned} \tag{41}$$

Denote $F_\theta'(x, \theta) = A(x, \theta) = (a_{i,j}(x, \theta))_{i,j}$. For T_1 we have

$$\begin{aligned} T_1 &= \sum_{i,j} \int_\delta^{1-\delta} (a_{i,j}(\hat{x}(t), \hat{\theta}_n) - a_{i,j}(x_\theta(t), \hat{\theta}_n))^2 w(t) dt \\ &= \sum_{i,j} \int_\delta^{1-\delta} \left(\int_0^1 \frac{\partial}{\partial x} a_{i,j}(x_\theta(t) + \lambda(\hat{x}(t) - x_\theta(t)), \hat{\theta}_n) d\lambda \, (\hat{x}(t) - x_\theta(t)) \right)^2 w(t) dt \\ &\leq \left(\sup_{t \in [\delta, 1-\delta]} \| \hat{x}(t) - x_\theta(t) \|^2 \right) \\ &\quad \times \sum_{i,j} \int_\delta^{1-\delta} \int_0^1 \left\| \frac{\partial}{\partial x} a_{i,j}(x_\theta(t) + \lambda(\hat{x}(t) - x_\theta(t)), \hat{\theta}_n) \right\|^2 d\lambda \, w(t) dt. \end{aligned}$$

By (11), as well as consistency of $\hat{\theta}_n$, Condition 4 and the continuous mapping theorem, the righthand side in the last inequality is of order

$$O_P(1) \left\{ \left(b^\alpha + \frac{1}{nb^2} \right)^2 + \frac{\log n}{nb} \right\}.$$

By a similar argument, the inequality

$$T_2 = \int_\delta^{1-\delta} \|F'_\theta(x_\theta(t),\hat{\theta}_n) - F'_\theta(x_\theta(t),\theta)\|^2 w(t)dt$$
$$\leq \|\hat{\theta}_n - \theta\|^2 \sum_{i,j} \int_\delta^{1-\delta} \int_0^1 \left\| \frac{\partial}{\partial\theta} a_{i,j}(x_\theta(t), \theta + \lambda(\hat{\theta}_n - \theta)) \right\|^2 d\lambda\, w(t)dt$$

holds. Here with some natural abuse of notation we first differentiate $a_{i,j}$ with respect to its second argument θ and only afterwards evaluate the obtained derivative at $x_\theta(t)$ and $\theta + \lambda(\hat{\theta}_n - \theta)$. Since the integrals at the righthand side of the above display are bounded in probability, we then get

$$\left\{ \int_\delta^{1-\delta} \|F'_\theta(x_\theta(t),\hat{\theta}_n) - F'_\theta(x_\theta(t),\theta)\|^2 w(t)dt \right\}^{1/2} = O_P(\|\hat{\theta}_n - \theta\|). \tag{42}$$

Now notice that (39) yields

$$\|\hat{\theta}_n - \theta\| \leq O_P(1) \Bigg(\left\| \int_\delta^{1-\delta} (F'_\theta(\hat{x}(t),\hat{\theta}_n))^T (\hat{x}'(t) - x'_\theta(t)) w(t)dt \right\|$$
$$+ \left\| \int_\delta^{1-\delta} (F'_\theta(\hat{x}(t),\hat{\theta}_n))^T (F(x_\theta(t),\theta) - F(\hat{x}(t),\theta)) w(t)dt \right\| \Bigg).$$

The Cauchy-Schwarz inequality then gives

$$\|\hat{\theta}_n - \theta\| \leq O_P(1) \left\{ \int_\delta^{1-\delta} \|F'_\theta(\hat{x}(t),\hat{\theta}_n)\|^2 w(t)dt \right\}^{1/2}$$
$$\times \left\{ \int_\delta^{1-\delta} \|\hat{x}'(t) - x'_\theta(t)\|^2 w(t)dt \right\}^{1/2}$$
$$+ O_P(1) \left\{ \int_\delta^{1-\delta} \|F'_\theta(\hat{x}(t),\hat{\theta}_n)\|^2 w(t)dt \right\}^{1/2}$$
$$\times \left\{ \int_\delta^{1-\delta} \|F(x_\theta(t),\theta) - F(\hat{x}(t),\theta)\|^2 w(t)dt \right\}^{1/2}.$$

By a by now standard argument, i.e. (11), (12), and the continuous mapping theorem, the righthand side can be further bounded to obtain

$$\|\hat{\theta}_n - \theta\| \leq O_P(1) \left(b^{\alpha-1} + \frac{1}{nb^3} + \sqrt{\frac{\log n}{nb^3}} + b^\alpha + \frac{1}{nb^2} + \sqrt{\frac{\log n}{nb}} \right). \tag{43}$$

Summarising the above results, we finally get that the second term at the righthand side of (40) satisfies

$$\left\| \int_\delta^{1-\delta} (F'_\theta(\hat{x}(t),\hat{\theta}_n) - F'_\theta(x_\theta(t),\theta))^T (\hat{x}'(t) - x'_\theta(t)) w(t)dt \right\|$$
$$\leq O_P(1) \left(b^{\alpha-1} + \frac{1}{nb^3} + \sqrt{\frac{\log n}{nb^3}} \right)^2 = o_P(n^{-1/2}),$$

where the last equality follows from our conditions on b. Here we also see that the condition $\alpha \geq 3$ is needed for the conclusion to hold.

To conclude the proof, it remains to consider the second term within brackets in (39). We have

$$\begin{aligned}
&\int_{\delta}^{1-\delta} (F'_\theta(\hat{x}(t),\hat{\theta}_n))^T (F(x_\theta(t),\theta) - F(\hat{x}(t),\theta))w(t)dt \\
&= \int_{\delta}^{1-\delta} (F'_\theta(x_\theta(t),\theta))^T (F(x_\theta(t),\theta) - F(\hat{x}(t),\theta))w(t)dt \qquad (44)\\
&\quad + \int_{\delta}^{1-\delta} (F'_\theta(\hat{x}(t),\hat{\theta}_n) - F'_\theta(x_\theta(t),\theta))^T (F(x_\theta(t),\theta) - F(\hat{x}(t),\theta))w(t)dt.
\end{aligned}$$

This can be analysed in a by now routine fashion, but we provide proofs. We first study the first term at the righthand side. By a standard argument we have

$$\begin{aligned}
&\int_{\delta}^{1-\delta} (F'_\theta(x_\theta(t),\theta))^T (F(x_\theta(t),\theta) - F(\hat{x}(t),\theta))w(t)dt \\
&= -\int_{\delta}^{1-\delta} (F'_\theta(x_\theta(t),\theta))^T \int_0^1 F'_x(x_\theta(t) + \lambda(\hat{x}(t) - x_\theta(t)),\theta)d\lambda\,(\hat{x}(t) - x_\theta(t))w(t)dt \\
&= -\int_{\delta}^{1-\delta} (F'_\theta(x_\theta(t),\theta))^T F'_x(x_\theta(t),\theta)(\hat{x}(t) - x_\theta(t))w(t)dt \\
&- \int_{\delta}^{1-\delta} (F'_\theta(x_\theta(t),\theta))^T \int_0^1 [F'_x(x_\theta(t)+\lambda(\hat{x}(t)-x_\theta(t)),\theta) - F'_x(x_\theta(t),\theta)]d\lambda\,(\hat{x}(t)-x_\theta(t))w(t)dt \\
&= T_3 + T_4.
\end{aligned}$$

Recalling (17), we see that T_3 appears in the leading term $\Gamma(\hat{x}) - \Gamma(x_\theta)$ in (16) and completes it together with the first term at the righthand side of (40). Next we consider T_4. Introduce the notation $F'_x(x,\theta) = B(x,\theta) = (b_{i,j}(x,\theta))_{i,j}$. We have

$$\begin{aligned}
&\left\| \int_0^1 [F'_x(x_\theta(t) + \lambda(\hat{x}(t) - x_\theta(t)),\theta) - F'_x(x_\theta(t),\theta)]d\lambda\,(\hat{x}(t) - x_\theta(t)) \right\| \\
&\leq \left(\sup_{t\in[\delta,1-\delta]} \|\hat{x}(t) - x_\theta(t)\| \right) \\
&\times \int_0^1 \|F'_x(x_\theta(t) + \lambda(\hat{x}(t) - x_\theta(t)),\theta) - F'_x(x_\theta(t),\theta)\| d\lambda \\
&\leq \left(\sup_{t\in[\delta,1-\delta]} \|\hat{x}(t) - x_\theta(t)\| \right) \\
&\times \int_0^1 \sum_{i,j} |b_{i,j}(x_\theta(t) + \lambda(\hat{x}(t) - x_\theta(t)),\theta) - b_{ij}(x_\theta(t),\theta)| d\lambda \\
&\leq \left(\sup_{t\in[\delta,1-\delta]} \|\hat{x}(t) - x_\theta(t)\| \right) \\
&\times \sum_{i,j} \int_0^1 \left\| \int_0^1 \frac{\partial}{\partial x} b_{ij}(x_\theta(t) + \kappa\lambda(\hat{x}(t) - x_\theta(t)),\theta) d\kappa \lambda(\hat{x}(t) - x_\theta(t)) \right\| d\lambda
\end{aligned}$$

$$\leq \left(\sup_{t \in [\delta, 1-\delta]} \|\hat{x}(t) - x_\theta(t)\|^2 \right)$$
$$\times \sum_{i,j} \int_0^1 \int_0^1 \left\| \frac{\partial}{\partial x} b_{ij}(x_\theta(t) + \kappa\lambda(\hat{x}(t) - x_\theta(t)), \theta) \right\| d\kappa d\lambda,$$

where in the last inequality we used the fact that $0 \leq \lambda \leq 1$. Since by convergence in probability of $\hat{x}$ to x_θ, Condition 4 and the continuous mapping theorem the integrals on the righthand side of the above display are bounded in probability, it follows from (11) that $\|T_4\|$ is

$$O_P(1) \left\{ \left(b^\alpha + \frac{1}{nb^2} \right)^2 + \frac{\log n}{nb} + \left(b^\alpha + \frac{1}{nb^3} \right) \sqrt{\frac{\log n}{nb}} \right\}.$$

This in turn is $o_P(n^{-1/2})$ because of the conditions on b. Finally, we treat the second term at the righthand side of (44). By the Cauchy-Schwarz inequality, its norm can be bounded by

$$\left\{ \int_\delta^{1-\delta} \|F'_\theta(\hat{x}(t), \hat{\theta}_n) - F'_\theta(x_\theta(t), \theta)\|^2 w(t) dt \right\}^{1/2}$$
$$\times \left\{ \int_\delta^{1-\delta} \|F(x_\theta(t), \theta) - F(\hat{x}(t), \theta)\|^2 w(t) dt \right\}^{1/2}.$$

Each of the terms at the righthand side have already been treated above, see (41) and (43), and it follows that the expression in the last display is $o_P(n^{-1/2})$. This concludes the proof of Proposition 3. □

Proof of Proposition 4. By a standard decomposition, we have

$$\begin{aligned} \mathbb{E}\left[(\Delta(\hat{\mu}_n) - \Delta(\mu))^2\right] &= (\mathbb{E}\left[\Delta(\hat{\mu}_n)\right] - \Delta(\mu))^2 + \mathbb{V}\mathrm{ar}\left[\Delta(\hat{\mu}_n)\right] \\ &= T_1^2 + T_2. \end{aligned}$$

The statement of the theorem will follow from Chebyshev's inequality, provided we show that the righthand side of the above display is $O\left(n^{-1}\right)$. For T_1 we have

$$\begin{aligned} |T_1| &= \left| \int_{\mathbb{R}} v(t) k(t) (\mathbb{E}\left[\hat{\mu}_n(t)\right] - \mu(t)) dt \right| \\ &\leq \sup_{t \in [\delta, 1-\delta]} \left|\mathbb{E}\left[\hat{\mu}_n(t)\right] - \mu(t)\right| \int_{\mathbb{R}} |v(t) k(t)| dt \\ &= O\left(b^\alpha + \frac{1}{nb^2} \right), \end{aligned}$$

where the last equality follows from (26). Taking $1/(2\alpha) \leq \gamma \leq 1/4$ gives that T_1 is $O\left(n^{-1/2}\right)$. We next consider T_2. By independence of the ϵ_i's, the fact that $\max_i |t_i - t_{i-1}| \lesssim n^{-1}$, boundedness of v and k, and integrability of K, we have

$$\begin{aligned} T_2 &= \mathbb{V}\mathrm{ar}\left[\sum_{i=1}^n (t_i - t_{i-1}) Y_i \int_\delta^{1-\delta} v(t) k(t) \frac{1}{b} K\left(\frac{t - t_i}{b} \right) dt \right] \\ &\lesssim \sigma^2 \sum_{i=1}^n (t_i - t_{i-1})^2 \left(\int_\delta^{1-\delta} v(t) k(t) \frac{1}{b} K\left(\frac{t - t_i}{b} \right) dt \right)^2 \end{aligned}$$

$$= O\left(\frac{1}{n}\right).$$

This completes the proof of Proposition 4. □

Proof of Theorem 1. The result is an easy consequence of Propositions 3 and 4. □

Appendix A

The proof of Proposition 1 is based on the following two lemmas, which provide integral approximations to the bias and variance of the estimator $\hat{\mu}_n$ and its derivative $\hat{\mu}_n'$ at a point t.

Lemma 1. *Let μ and K be continuously differentiable and let K be supported on the interval $[-1,1]$. For any $t \in [0,1]$*

$$\mathbb{E}\left[\hat{\mu}_n(t)\right] = \int_0^1 \mu(s)\frac{1}{b}K\left(\frac{t-s}{b}\right)ds + O\left(\frac{1}{nb^2}\right) \tag{45}$$

holds in the regression model (7)*. The order bound on the remainder term in* (45) *is uniform in $t \in [0,1]$.*

Proof. The proof is based on the Riemann sum approximation of the integral. Since $\mathbb{E}\left[\epsilon_i\right] = 0$, we have

$$\begin{aligned}\mathbb{E}\left[\hat{\mu}_n(t)\right] &= \int_0^1 \mu(s)\frac{1}{b}K\left(\frac{t-s}{b}\right)ds \\ &\quad - \int_0^1 \mu(s)\frac{1}{b}K\left(\frac{t-s}{b}\right)ds + \sum_{i=1}^n (t_i - t_{i-1})\mu(t_i)\frac{1}{b}K\left(\frac{t-t_i}{b}\right).\end{aligned}$$

The first term at the righthand side of this expression is the first term of (45). We will now establish an upper bound on the difference of the other two terms. Using continuous differentiability of μ and K and the fact that $\max_i |t_i - t_{i-1}| = O(n^{-1})$, we have

$$\begin{aligned}&\left|\int_0^1 \mu(s)\frac{1}{b}K\left(\frac{t-s}{b}\right)ds - \sum_{i=1}^n (t_i - t_{i-1})\mu(t_i)\frac{1}{b}K\left(\frac{t-t_i}{b}\right)\right| \\ &\quad = \left|\sum_{i=1}^n \int_{t_{i-1}}^{t_i}\left\{\mu(s)\frac{1}{b}K\left(\frac{t-s}{b}\right) - \mu(t_i)\frac{1}{b}K\left(\frac{t-t_i}{b}\right)\right\}ds\right| \\ &\quad \leq \sum_{i=1}^n \int_{t_{i-1}}^{t_i}\left|\mu(s)\frac{1}{b}K\left(\frac{t-s}{b}\right) - \mu(s)\frac{1}{b}K\left(\frac{t-t_i}{b}\right)\right|ds \\ &\quad + \sum_{i=1}^n \int_{t_{i-1}}^{t_i}\left|\mu(s)\frac{1}{b}K\left(\frac{t-t_i}{b}\right) - \mu(t_i)\frac{1}{b}K\left(\frac{t-t_i}{b}\right)\right|ds \\ &\qquad\qquad \lesssim \frac{1}{nb^2}\|\mu\|_\infty\|K'\|_\infty + \frac{1}{nb}\|\mu'\|_\infty\|K\|_\infty,\end{aligned}$$

which is of order $n^{-1}b^{-2}$. This establishes (45). □

The second lemma can be proved along the same lines as the previous one and therefore we omit its proof. The existence of the second derivative of K is needed in the proof of this lemma.

Lemma 2. *Let μ be continuously differentiable and let K be twice continuously differentiable and be supported on the interval $[-1,1]$. For all $t \in [0,1]$*

$$\mathbb{E}\left[\hat{\mu}_n'(t)\right] = \int_0^1 \mu(s)\frac{1}{b^2}K'\left(\frac{t-s}{b}\right)ds + O\left(\frac{1}{nb^3}\right) \tag{46}$$

holds in the regression model (7)*. Furthermore, if $b \le \delta$ and $t \in [\delta, 1-\delta]$, then integration by parts yields*

$$\mathbb{E}\left[\hat{\mu}_n'(t)\right] = \int_{-1}^1 \mu'(t-bu)K(u)du + O\left(\frac{1}{nb^3}\right). \tag{47}$$

The order bounds on the remainder terms in (46) *and* (47) *are uniform in t.*

The following lemma is used in the proof of Proposition 2.

Lemma 3. *Let the stochastic process $X_n = (X_{n,\eta})_{\eta\in\Theta}$ be defined as*

$$X_n = (X_{n,\eta})_{\eta\in\Theta} = \left(\int_\delta^{1-\delta} \|F(\hat{x}(t),\eta) - F(x_\theta(t),\eta)\|^2 w(t)dt\right)_{\eta\in\Theta}.$$

Then under the conditions of Proposition 2 we have $X_n \overset{\mathrm{P}}{\to} 0$, where 0 at the right-hand side denotes the zero process on Θ and convergence is understood as convergence for random elements with values in the space $C(\Theta)$ of continuous functions on Θ, which is equipped with the supremum norm.

Proof. To prove the lemma, we will verify the conditions of Theorem 18.14 of van der Vaart (1998). By (11) and the continuous mapping theorem, see Theorem 18.11 in van der Vaart (1998), for every fixed η it holds that

$$\int_\delta^{1-\delta} \|F(\hat{x}(t),\eta) - F(x_\theta(t),\eta)\|^2 w(t)dt \overset{\mathrm{P}}{\to} 0. \tag{48}$$

Consequently, for any positive integer k and any $\eta_1,\ldots,\eta_k \in \Theta$ we have

$$(X_{n,\eta_1},\ldots,X_{n,\eta_k}) \rightsquigarrow (\underbrace{0,\ldots,0}_{k})$$

and hence condition (i) of Theorem 18.14 in van der Vaart (1998) is satisfied. Introduce

$$G = \bigcap_{j=1}^d \left\{\sup_{t\in[\delta,1-\delta]} |\hat{x}_j(t) - x_{\theta j}(t)| \le \beta\right\}$$

and notice

$$G^c = \bigcup_{j=1}^d \left\{\sup_{t\in[\delta,1-\delta]} |\hat{x}_j(t) - x_{\theta j}(t)| > \beta\right\}.$$

For any positive ε and β and any partition $\Theta_1,\ldots,\Theta_m$ of Θ we have

$$\begin{aligned} &P\left(\sup_\ell \sup_{\eta,\zeta\in\Theta_\ell} |X_{n,\eta} - X_{n,\zeta}| \ge \varepsilon\right) \\ &\quad\le P\left(\sup_\ell \sup_{\eta,\zeta\in\Theta_\ell} |X_{n,\eta} - X_{n,\zeta}| \ge \varepsilon; G\right) + P\left(G^c\right). \end{aligned} \tag{49}$$

By (11) we know that

$$\lim_{n\to\infty} P\left(G^c\right) \leq \lim_{n\to\infty} \sum_{j=1}^{d} P\left(\sup_{t\in[\delta,1-\delta]} |\hat{x}_j(t) - x_{\theta j}(t)| > \beta \right) = 0. \tag{50}$$

We will now show that for arbitrarily small positive ρ and ε there exists a partition $\Theta_1, \ldots, \Theta_m$ of Θ, such that

$$\limsup_{n\to\infty} P\left(\sup_{\ell} \sup_{\eta,\zeta\in\Theta_\ell} |X_{n,\eta} - X_{n,\zeta}| \geq \varepsilon; G \right) \leq \rho.$$

Together with (49) and (50) this will imply condition (ii) of Theorem 18.14 in van der Vaart (1998) and hence also the fact that X_n converges weakly to zero. The statement of the lemma will then be a simple consequence of the fact that convergence in distribution and in probability are equivalent for constants, see Theorem 18.10 of van der Vaart (1998).

Notice that

$$\begin{aligned}
&|X_{n,\eta} - X_{n,\zeta}| \\
&\qquad \leq \int_\delta^{1-\delta} \|F(\hat{x}(t),\eta) - F(x_\theta(t),\eta) - F(\hat{x}(t),\zeta) + F(x_\theta(t),\zeta)\| \\
&\qquad\quad \times (\|F(\hat{x}(t),\eta) - F(x_\theta(t),\eta)\| + \|F(\hat{x}(t),\zeta) - F(x_\theta(t),\zeta)\|) w(t)dt \\
&\quad \leq \left\{ \int_\delta^{1-\delta} \|F(\hat{x}(t),\eta) - F(x_\theta(t),\eta) - F(\hat{x}(t),\zeta) + F(x_\theta(t),\zeta)\|^2 w(t)dt \right\}^{1/2} \\
&\quad \times \left\{ \int_\delta^{1-\delta} (\|F(\hat{x}(t),\eta) - F(x_\theta(t),\eta)\| + \|F(\hat{x}(t),\zeta) - F(x_\theta(t),\zeta)\|)^2 w(t)dt \right\}^{1/2} \\
&\quad = \sqrt{T_3}\sqrt{T_4}.
\end{aligned}$$

For T_3 we have

$$\begin{aligned}
T_3 \leq 2 &\int_\delta^{1-\delta} \|F(\hat{x}(t),\eta) - F(\hat{x}(t),\zeta)\|^2 w(t)dt \\
&\qquad\qquad + 2\int_\delta^{1-\delta} \|F(x_\theta(t),\eta) - F(x_\theta(t),\zeta)\|^2 w(t)dt.
\end{aligned}$$

Restricting ω's from the sample space Ω to the set G, we get

$$\begin{aligned}
T_3 &\leq 2\int_\delta^{1-\delta} \int_0^1 \|F_\theta'(\hat{x}(t),\zeta + \lambda(\eta-\zeta))\|^2 d\lambda \, \|\eta-\zeta\|^2 w(t)dt \\
&\quad + 2\int_\delta^{1-\delta} \int_0^1 \|F_\theta'(x_\theta(t),\zeta + \lambda(\eta-\zeta))\|^2 d\lambda \, \|\eta-\zeta\|^2 w(t)dt \\
&\leq 4\|\eta-\zeta\|^2 \int_\delta^{1-\delta} w(t)dt \sup_{\substack{\|x_j\|\leq\|x_{\theta j}\|_\infty+\beta, j=1,\ldots,d \\ \nu\in\Theta}} \|F_\theta'(x,\nu)\| = C(\beta,w,\theta,\Theta)\|\eta-\zeta\|^2
\end{aligned}$$

on the set G. Notice that $C(\beta,w,\theta,\Theta)$ is a finite constant, because $\|F_\theta'(x,\nu)\|$ is continuous and its supremum is taken over a compact set. By similar techniques one can show that $T_4 \leq C'(\beta, w, \theta, \Theta)$ for some constant $C'(\beta, w, \theta, \Theta)$ which depends only on β, w, θ, and Θ. Consequently,

$$\begin{aligned} &P\left(\sup_{\ell} \sup_{\eta,\zeta\in\Theta_\ell} |X_{n,\eta} - X_{n,\zeta}| \geq \varepsilon; G\right) \\ &\qquad \leq P\left(\sup_{\ell} \sup_{\eta,\zeta\in\Theta_\ell} \sqrt{C(\beta,w,\theta,\Theta)C'(\beta,w,\theta,\Theta)}\,\|\eta-\zeta\| \geq \varepsilon\right). \end{aligned} \tag{51}$$

Now take a partition $\Theta_1, \ldots, \Theta_m$ of Θ such that for all $\ell = 1, \ldots, m$

$$0 < \operatorname{diam}\Theta_\ell < \frac{\varepsilon}{\sqrt{C(\beta,w,\theta,\Theta)C'(\beta,w,\theta,\Theta)}}$$

holds, where $\operatorname{diam}\Theta_\ell$ denotes the diameter of the set Θ_ℓ. Observe that since $\Theta \subset \mathbb{R}^p$ is compact, there indeed exists a finite m for which this is satisfied. The righthand side of (51) for such a partition is zero and consequently the conditions (i) and (ii) of Theorem 18.14 of van der Vaart (1998) hold. This completes the proof of the lemma. □

Appendix B

Here we state and prove a modification of Proposition 1 for the case when the ϵ_i's are bounded.

Proposition 5. *In the regression model* (7) *replace the assumption of Gaussianity of the ϵ_i's by $|\epsilon_i| \leq C$ for some constant $C > 0$ and suppose Condition 5 holds.*

(i) If μ is $\alpha \geq 1$ times continuously differentiable and $b \to 0$ as $n \to \infty$, then

$$\sup_{t\in[\delta,1-\delta]} |\hat{\mu}_n(t) - \mu(t)| = O_P\left(b^\alpha + \frac{1}{nb^2} + \sqrt{\frac{\log n}{nb}}\right). \tag{52}$$

(ii) If μ is $\alpha \geq 2$ times continuously differentiable and $b \to 0$ as $n \to \infty$, then

$$\sup_{t\in[\delta,1-\delta]} |\hat{\mu}_n'(t) - \mu'(t)| = O_P\left(b^{\alpha-1} + \frac{1}{nb^3} + \sqrt{\frac{\log n}{nb^3}}\right) \tag{53}$$

is valid. Moreover, $\hat{\mu}_n$ and $\hat{\mu}_n'$ are consistent on $[\delta, 1-\delta]$, if $nb^3/\log n \to \infty$ holds additionally.

Proof. The proof of (52) follows the same steps as the proof of (9). The only difference is that we need to show that

$$\mathbb{E}\left[\max_{1\leq j\leq N} |Z_j|^2\right] = O\left(\frac{\log n}{nb}\right) \tag{54}$$

holds also for bounded ϵ_i's and not only for the Gaussian ϵ_i's. To this end we will use some results from Chapter 2.2 of Wellner and van der Vaart (1996). Let η be a nondecreasing and convex function on $[0, \infty)$, such that $\eta(0) = 0$. The Orlicz norm $\|X\|_\eta$ of a random variable X is defined as

$$\|X\|_\eta = \inf\left\{C > 0 : \mathbb{E}\left[\eta\left(\frac{|X|}{C}\right)\right] \leq 1\right\}.$$

A particular η that we will use is $\eta(x) = \exp(x^2) - 1$. Since the ϵ_i's have mean zero and are bounded, for any $x > 0$ Hoeffding's inequality, see Theorem 2 in Hoeffding (1963), implies

$$P(|Z_j| > x) \leq 2 \exp\left(-2x^2 / \left(\sum_{i=1}^{n} C^2 (S_i(s_j))^2 \right) \right).$$

By Condition 1

$$\begin{aligned} C^2 \sum_{i=1}^{n} (S_i(s_j))^2 &\lesssim C^2 \|K\|_\infty^2 \frac{1}{n^2 b^2} \sum_{i=1}^{n} 1_{[|s_j - t_i| \leq b]} \\ &\leq \frac{1}{nb} C^2 \|K\|_\infty^2 c_1 \max\left(2, \max_n \frac{1}{nb}\right) = \frac{1}{C_0 nb} \end{aligned}$$

holds. Thus the inequality

$$P(|Z_j| > x) \leq 2 \exp(-2C_0 nbx^2)$$

is valid. By Lemma 2.2.1 of Wellner and van der Vaart (1996) it then follows that

$$\max_j \|Z_j\|_\eta \leq \frac{C_1}{\sqrt{nb}}, \tag{55}$$

where C_1 depends on C_0 only. Let $\|X\|_2$ denote the L_2 norm of a random variable X, i.e. $\|X\|_2 = \sqrt{\mathbb{E}\left[X^2\right]}$. Notice that the inequality

$$\|X\|_2 \leq \|X\|_\eta, \tag{56}$$

holds, because of $\eta(x) \geq x^2$. The inequalities (55) and (56) combined with Lemma 2.2.2 of Wellner and van der Vaart (1996) yield that

$$\sqrt{\mathbb{E}\left[\max_{1 \leq j \leq N} |Z_j|^2\right]} \leq \frac{C_3}{\sqrt{nb}} \eta^{-1}(N),$$

where the constant C_3 is independent of N. Now notice that for $N \geq 4$

$$\eta^{-1}(N) = \sqrt{\log(N+1)} \leq \sqrt{\log(N^2)} = 2\sqrt{\log n}.$$

Hence (54) holds and this completes the proof of (52). Formula (53) can be proved in a similar fashion. □

Acknowledgements The research reported here was started when the first author was a Postdoc at EURANDOM, Eindhoven, and the second one was a Senior Fellow there.

References

V.I. Arnold. *Ordinary Differential Equations.* MIT Press, Massachusets, 1973.

R. Bellman and R.S. Roth. The use of splines with unknown end points in the identification of systems. *J. Math. Anal. Appl.*, 34:2633, 1971.

J.K. Benedetti. On the nonparametric estimation of regression functions. *J. Roy. Statist. Soc. Ser. B*, 39:248–253, 1977.

P.J. Bickel and Y. Ritov. Nonparametric estimators which can be "plugged-in". *Ann. Statist.*, 31:1033–1053, 2003.

P.J. Bickel, C.A.J. Klaassen, Y. Ritov and J.A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models.* Springer, New York, 1998.

H.G. Bock. Recent advances in parameter identification techniques for O.D.E. In P. Deurfland and E. Hairer, editors, *Numerical Treatment of Inverse Problems in Differential and Integral Equations*. Birkhäuser, Boston, 95–121, 1983.

N.J-B. Brunel. Parameter estimation of ODE's via nonparametric estimators. *Electron. J. Stat.*, 2:1242–1267, 2008.

I-C. Chou and E.O. Voit. Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Math. Biosci.*, 219:57–83, 2009.

L. Edelstein-Keshet. *Mathematical Models in Biology.* Society for Industrial and Applied Mathematics, Philadelphia, 2005.

S.P. Ellner, Y. Seifu and R.H. Smith. Fitting population dynamic models to time-series data by gradient matching. *Ecology*, 83:2256-2270, 2002.

A.J. van Es. *Aspects of Nonparametric Density Estimation.* CWI Tract, 77. Stichting Mathematisch Centrum, Centrum voor Wiskunde en Informatica, Amsterdam, 1991.

J. Fan and J.S. Marron. Fast implementations of nonparametric curve estimators. *J. Comput. Graph. Stat.*, 3:35-56, 1994.

M. Feinberg. *Lectures on Chemical Reaction Networks.* Lectures delivered at the Mathematics Research Center, University of Wisconsin-Madison, 1979.

Th. Gasser, H-G. Müller and V. Mammitzsch. Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B*, 47:238–252, 1985.

Th. Gasser and H-G. Müller. Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.*, 11:197–211, 1984.

S. van de Geer. Estimating a regression function. *Ann. Statist.*, 18:907-924, 1990.

S. van de Geer and M. Wegkamp. Consistency for the least squares estimator in nonparametric regression. *Ann. Statist.*, 24:2513-2523, 1996.

A. Gelman, F.Y. Bois and J. Jiang. Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *J. Amer. Statist. Assoc.*, 91:1400–1412, 1996.

M. Girolami. Bayesian inference for differential equations. *Theor. Comput. Sci.*, 408:4–16, 2008.

L. Goldstein and K. Messer. Optimal plug-in estimators for nonparametric functional estimation. *Ann. Statist.*, 20:1306-1328, 1992.

E. Hairer and G. Wanner. *Solving Ordinary Differential Equations. II. Stiff and Differential-Algebraic Problems.* Springer-Verlag, Berlin, 1996.

P. Hall and J.S. Marron. On variance estimation in nonparametric regression. *Biometrika*, 77:415–419, 1990.

P.W. Hemker. Numerical methods for differential equations in system simulation and in parameter estimation. In H.C. Hemker and B. Hess, editors, *Analysis and Simulation of Biochemical Systems*. North Holland Publ. Comp., Amsterdam, 59–80, 1972.

W.S. Hlavacek and M.A. Savageau. Rules for coupled expression of regulator and effector genes in inducible circuits. *J. Mol. Biol.*, 255:121–139, 1996.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13-30, 1963.

G. Hooker. Forcing function diagnostics for nonlinear dynamics. *Biometrics*, 65:928–936, 2009.

P.J. Huber. *Robust Statistics.* John Wiley & Sons, Inc., New York, 1981.

R.I. Jennrich. Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.*, 40:633–643, 1969.

M.C. Jones, J.S. Marron and S.J. Sheather. A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.*, 91:401–407, 1996.

S. Kikuchi, D. Tominaga, M. Arita, K. Takahashi and M. Tomita. Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics*, 19:643–650, 2003.

H. Liang and H. Wu. Parameter estimation for differential equation models using a framework of measurement error in regression models. *J. Amer. Statist. Assoc.*, 103:1570–1583, 2008.

C.R. Loader. Bandwidth selection: classical or plug-in? *Ann. Statist.*, 27:415–438, 1999.

D.W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM J. Appl. Math.*, 11:431–441, 1963.

T.L. McMurry and D.N. Politis. Nonparametric regression with infinite order flat-top kernels. *J. Nonparametr. Stat.*, 16:549–562, 2004.

K. Messer and L. Goldstein. A new class of kernels for nonparametric curve estimation. *Ann. Statist.*, 21:179–195, 1993.

D. Pollard and P. Radchenko. Nonlinear least-squares estimation. *J. Multivariate Anal.*, 97:548–562, 2006.

M.B. Priestley and M.T. Chao. Non-parametric function fitting. *J. Roy. Statist. Soc. Ser. B*, 34:385–392, 1972.

X. Qi and H. Zhao. Asymptotic efficiency and finite sample properties of the generalized profiling estimation of parameters in ordinary differential equations. *Ann. Statist.*, 38:435–481, 2010.

J.O. Ramsay, G. Hooker, D. Campbell and J. Cao. Parameter estimation for differential equations: a generalized smoothing approach. With discussions and a reply by the authors. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69:741–796, 2007.

E. Schuster and S. Yakowitz. Contributions to the theory of nonparametric regression, with application to system identification. *Ann. Statist.*, 7:139–149, 1979.

E.D. Sontag. Structure and stability of certain chemical networks and applications to the kinetic proofreading model of T-cell receptor signal transduction. *IEEE Trans. Automat. Control*, 46:1028–1047, 2001.

S.M. Stigler. Gauss and the invention of least squares. *Ann. Statist.*, 9:465–474, 1981.

W.J.H. Stortelder. Parameter estimation in dynamic systems. *Math. Comput. Simulat.*, 42:135–142, 1996.

G. Szegö. *Orthogonal Polynomials.* American Mathematical Society, Providence, 1975.

A.B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, New York, 2009.

A.W. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, Cambridge, 1998.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer-Verlag, New York, 1996.

J.M. Varah. A spline least squares method for numerical parameter estimation in differential equations. *SIAM J. Sci. Stat. Comput.*, 3:28–46, 1982.

E.O. Voit. *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists.* Cambridge University Press, Cambridge, 2000.

E.O. Voit and J. Almeida. Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics*, 10:1670–1681, 2004.

E.O. Voit and M.A. Savageau. Power-law approach to modeling biological systems; III. Methods of analysis. *J. Ferment. Technol.*, 60:233-241, 1982.

A. Wächter and L.T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Program. Ser. A*, 106:25–57, 2006.

M.P. Wand and W. Schucany. Gaussian-based kernels. *Can. J. Stat.* 18:197–204, 1990.

M.P. Wand and M.C. Jones. *Kernel Smoothing.* Chapman and Hall, London, 1995.

L. Wasserman. *All of Nonparametric Statistics.* Springer, New York, 2006.

Wolfram Research, Inc. Mathematica, Version 6.0. Champaign, IL, 2007.

C-F. Wu. Asymptotic theory of nonlinear least squares estimation. *Ann. Statist.*, 9:501-513, 1981.

H. Xue, H. Miao and H. Wu. Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error. *Ann. Statist.*, 38:2351-2387, 2010.

Department of Mathematics, VU University Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
E-mail address: s.gugushvili@vu.nl

Korteweg-de Vries Institute for Mathematics, Universiteit van Amsterdam, P.O. Box 94248, 1090 GE Amsterdam, The Netherlands
E-mail address: C.A.J.Klaassen@uva.nl